Facilitating the Compilation and Dissemination of Ad-Hoc Web Corpora

William H. Fletcher
United States Naval Academy,[1] USA

Final revision, 29 April2004. To appear in:
Aston, Guy, Silvia Bernardini and Dominic Stewart (2004*). Papers from the Fifth International Conference on Teaching and Language Corpora.* Amsterdam: Benjamins.

Current version online at
http://kwicfinder.com/Facilitating_Compilation_and_Dissemination_of_Ad-Hoc_Web_Corpora.pdf

---

Since the World Wide Web gained prominence in the mid-1990s it has tantalized language investigators and instructors as a virtually unlimited source of machine-readable texts for compiling corpora and developing teaching materials. The broad range of languages and content domains found online also offers translators enormous promise both for translation-by-example and as a comprehensive supplement to published reference works. This paper surveys the impediments which still prevent the Web from realizing its full potential as a linguistic resource and discusses tools to overcome the remaining hurdles. Identifying online documents which are both relevant and reliable presents a major challenge. As a partial solution the author's Web concordancer KWiCFinder automates the process of seeking and retrieving webpages. Enhancements which permit more focused queries than existing search engines and provide search results in an interactive exploratory environment are described in detail. Despite the efficiency of automated downloading and excerpting, selecting Web documents still entails significant time and effort. To multiply the benefits of a search, an online forum for sharing annotated search reports and linguistically interesting texts with other users is outlined. Furthermore, the orientation of commercial search engines toward the general public makes them less beneficial for linguistic research. The author sketches plans for a specialized Search Engine for Applied Linguists and a selective Web Corpus Archive which build on his experience with KWiCFinder. He compares his available and proposed solutions to existing resources, and surveys ways to exploit them in language teaching. Together these proposed services will enable language learners and professionals to tap into the Web effectively and efficiently for instruction, research and translation.

1. Aperitivo

Aston (2002) compares learner-compiled corpora to professionally produced corpora through a memorable analogy to fruit salad. While home-made fruit salad (and corpora) can entail various benefits he enumerates, the off-the-shelf variety offers reliability and convenience, supplemented in its corpus analogue by documentation and specialized software. He proposes that learners can follow a compromise "pick'n'mix" strategy, compiling their own customized subcorpora from professionally selected materials.

---

[1] Research for this paper was supported in part by the Naval Academy Research Council.

By now this alimentary analogy (but by no means the strategy) must have passed its "best-by" date, yet I cannot resist adapting it to the World Wide Web. Food-borne analogies seem very appropriate for a conference in Bertinoro, the historic town of culinary and oenological hospitality, so I begin and end on this note.

For years the Web has tantalized language professionals, offering a boundless pool of texts whose fruitful exploitation has remained out of reach. It is like an old-fashioned American community pot-luck supper, to which each family brings a dish to share with the other guests. As a child at such events I would taste many dishes in search of the most flavorful; usually I wasted my appetite sampling mediocre fare. Similarly I have spent countless hours online seeking and sifting through webpages, too often squandering my time, then giving up, sated yet unsatisfied.

Frustration with finding useful content in the World Wide Haystack inspired me to design and implement the Web concordancing tools and strategies described here which enable users to compile ad-hoc corpora from webpages.[2] Reflection on essential needs unmet by this model has led me to chart the course for future development to make sharing of Web corpora easier and more rewarding, and to outline an infrastructure for a search engine tailored to the needs of language professionals and learners. My conviction is simple: if online linguistic research can be made effective and efficient, linguists and learners will not have to take pot-luck with what they find on the Web by chance.

2. Web as corpus ? !

A haphazard accumulation of machine-readable texts, the World Wide Web is unparalleled for quantity, diversity and topicality. This ever-expanding body of documents now encompasses at least 10 billion (10^9) webpages publicly available via links, with several times that number in the "hidden" Web accessible only through database queries or passwords. Once overwhelmingly Anglophone, the Web now encompasses languages used by a majority of the world's population. Currently native English speakers account for only 35% of Web users, and their relative prominence is dwindling as the Web expands into more non-western language areas.[3] Online content covers virtually every knowledge domain of interest to language professionals or learners, and incorporates contemporary issues and emerging usage rare in customary sources.

With all the Web offers, why have all but a handful of corpus linguists and language professionals failed to exploit this vast potential source for corpora?[4] Surely the effort required to locate relevant, reliable documents outweighs all other explanations for this neglect. The quantity of information online greatly surpasses its overall quality. Unpolished ephemera abound alongside rare treasures, and online documents generally seem to consist more of accumulations of fragments, stock phrases and bulleted lists than of original extended text. Among the longer

---

[2] Ad-hoc corpora – also designated as "disposable" or "do-it-yourself" corpora – are compiled to meet a specific information need and typically abandoned once that need has been met (see e.g., Varantola 2003 and Zanettin 2001).

[3] Figures from September 2003 (http://www.global-reach.biz/globstats/, visited 26 February 2004), which estimates the online population of native speakers of English and of other European languages at 35% each, while speakers of other languages total about 30%. These number contrast sharply with the late 1990s, when English speakers comprised over three-quarters of the world's online population.

[4] The number of linguists exploiting the Web as a linguistic corpus (beyond the casual "let's see how many hits I can find for this on Google") is growing. Kilgarriff and Grefenstette (2003) survey numerous papers and projects in this field. Other representative examples of applying Web data to specific linguistic problems include Banko and Brill (2001), Grefenstette (1999), and Volk (2002). Brekke (2002) and Fletcher (2001 a, b) discuss the pitfalls and limitations of the Web as a corpus. Finally, researchers like De Schryver (2002), Ghani et al. (2001) and Scannell (2004) demonstrate the importance of the Web for compiling corpora of minority languages for which other electronic and even print sources are severely limited.

coherent texts, specialized genres such as commercial, journalistic, administrative and academic documents predominate. Assessing the "authoritativeness" of a webpage–the accuracy of its content and representativeness of its linguistic form–demands time and expertise.

Despite these challenges, there are compelling reasons to supplement existing corpora with online materials. A static corpus represents a snapshot of issues and language usage known when it was compiled. The great expense of setting up a large corpus precludes frequent supplementation or replacement, and contemporary content can grow stale quickly. In contrast, new documents appear on the Web daily, so up-to-date content and usage tend to be well represented online. In addition, even a very large corpus might include few examples of infrequent expressions or constructions that can be found in abundance on the Web. Moreover, certain content domains or text genres may be underrepresented in an existing corpus or even missing entirely. With the Web as a source one usually can locate documents from which to compile an ad-hoc corpus to meet the specific needs of groups of investigators, translators or learners. Finally, while existing corpora may entail significant fees and require specialized hardware and software to consult, Web access is generally inexpensive, and desktop computers to perform the necessary processing are now within the reach of students as well as researchers.

3. Locating forms and content on line

3.1 Established techniques

Marcia Bates' "information search tactics" can be adapted to categorize typical approaches to finding useful material online (Fletcher 2001b). *Hunting,* or searching directly for specific forms or content online, appears to be the most widely-used and productive strategy. For specialized content, *grazing,* i.e., focusing on predetermined reliable websites, has also proved an effective strategy for corpus construction.[5] In contrast to these goal-oriented tactics, *browsing,* the archetypal Web activity, relies on serendipity for the user to discover relevant material. What follows shows how all three strategies can be improved to make the Web a more accessible corpus for language research and learning.

"Hunting" via Web searches is the most effective means of locating online content. Unfortunately this strategy depends on commercial search engines and thus is limited by their quirks and weaknesses. A dozen main search engines aspire to "crawl" and map the entire Web, yet none indexes more than roughly a fifth of the publicly-accessible webpages. Thousands of specialized search engines focus on narrower linguistic, geographic or knowledge domains. The search process is familiar to all Web users: first one formulates a query to find webpages with specific words or phrases and submits it to a search engine. Some search engines support "smart features" for a few major languages, for example to search automatically for synonyms or alternate word forms ("stemming"). Meta-search engines query several search engines simultaneously, then "collapse" the results into a single list of unique links. In all cases, however, the user must still retrieve and evaluate the documents individually for relevance and reliability.

Beyond the tedium of winnowing the wheat from the chaff, this search-and-select strategy has several flaws, starting with the port-of-entry to the Web. Search engines are not research libraries but commercial enterprises targeted at the needs of the general public. The availability and

[5] Knut Hofland's Norwegian newspaper corpus (Hofland 2002) follows a "grazing" strategy to "harvest" articles daily from several newspapers. Using material from a limited number of sites offers several advantages: permission and cooperation for use of texts can be secured; recurring page layouts help distinguish novel content from "boilerplate" materials automatically; the texts' genre and content domain are predictable, and their authorship, representativeness and reliability can be established. Similarly, GlossaNet (http://glossa.fltr.ucl.ac.be), described in greater detail below, monitors and analyzes text from over 100 newspapers in nine languages, but does not archive them for public access.

implementation of their services change constantly: features are added or dropped to mimic or outdo the competition; acquisitions and mergers threaten their independence; financial uncertainties and legal battles challenge their very survival. The search sites' quest for revenue can diminish the objectivity of their search results, and various "page ranking" algorithms may lead to results that are not representative of the Web as a whole.[6] Most frustrating is the minimal support for the requirements of serious researchers: current trends lead away from sites like AltaVista supporting sophisticated complex queries (which few users employ) to ones like Google offering only simple search criteria. In short, the search engines' services are useful to investigators by coincidence, not design, and researchers are tolerated on mainstream search sites only as long as their use does not affect site performance adversely.

3.2 KWiCFinder Web concordancer

To overcome some limitations of general-purpose search engines and to automate aspects of the process of searching and selecting I have developed the search agent *KWiCFinder,* short for *Key Word in Context Finder.* This free research tool[7] helps users create a well-formed query and submits it to the AltaVista search engine. It then retrieves and produces a KWiC concordance of 5–15 online documents per minute without further attention from the user; dead links and documents whose content no longer matches the query are excluded from this search report. Here I discuss how it enhances the search process for language analysis as background to the proposals advanced in the *solutions* sections below; for greater detail see the website referenced in the previous note and Fletcher 2001b.

3.2.1 Searching with KWiCFinder

To streamline the document selection process, KWiCFinder features more narrowly focused search criteria than commercial search sites. For example, AltaVista supports the wildcard **,* which stands for any sequence of zero to five characters. KWiCFinder adds the wildcards *?* and *%,* which represent "exactly one" and "zero or one" characters respectively. In an AltaVista query, lower-case and "plain" characters match their upper-case and accented counterparts, so that e.g., *a* in a query would match any of *aáâäàãæåAÁÂÄÀÃÆÅ.* KWiCFinder introduces the "sic" option, which forces an *exact* match of lower-case and "plain" characters. For example, choosing "sic" distinguishes the past tense of the German passive auxiliary *wurde* from the subjunctive auxiliary *würde,* and both are kept separate from the noun *Würde* "dignity". Similarly, KWiCFinder supports the operators BEFORE and AFTER in addition to AltaVista's NEAR to relate multiple search terms, and permits the user to specify how many words may separate them. These enhancements do come at a price: KWiCFinder must submit a standard query to AltaVista and retrieve all matching documents, then filter out webpages not meeting the narrower search criteria. In extreme cases, dozens of webpages must be downloaded and analyzed to find one that matches the searcher's query exactly.

Obviously the most efficient searches forgo wildcards by specifying and matching variant forms exactly. Especially in morphologically rich languages, entering all possible variants into a query can be most tedious. KWiCFinder introduces three types of "tamecards," a shorthand notation for such variants. A simple tamecard pattern is entered between [ ], with variants separated by commas: *work[,s,ed,ing]* is expanded to *work* OR *works* OR *worked* OR *working*, but it does not match *worker, workers,* as would wildcard *work*.* Indexed tamecards appear between { }; each variant

---

[6] "Paid positioning" and other "revenue-stream enhancers" may put advertisers' webpages at the top of the search results. The link popularity ranking strategy exemplified by Google—webpages to which more other sites link are ranked before relatively unknown pages—can mask much of the Web's diversity by favoring well-known sites.

[7] KWiCFinder is available free online from http://KWiCFinder.com (alternate URL http://miniappolis.com/KWiCFinder). First demonstrated at CALICO 1999 and available online since later that year, it is described in far greater detail in Fletcher 2001b.

is combined with the corresponding variant in other indexed tamecards in the same query field. For instance, *{me,te,se} lav{o,as,a}* expands only to the Spanish reflexive forms *me lavo, te lavas, se lava,* but not to non-reflexive *te lavo* or ungrammatical **se lavo.*

KWiCFinder's "implicit tamecards" with hyphen or apostrophe match forms both with and without the punctuation mark and / or space: *on-line* matches the common variants *on line, online, on-line.* This is particularly useful for English, with its great variation in writing compounds with and without spaces and hyphens, and for German, where the new spelling puts asunder forms that formerly were joined*: kennen-lernen* matches both traditional *kennenlernen* and reformed *kennen lernen*, which coexist in current practice and are reunited in a KWiCFinder search. As a final implicit set of tamecards KWiCFinder recognizes the equivalence of some language-specific orthographic variants, such as German *ß* and *ss, ä ö ü* and *ae oe ue.*

3.2.2 Exploring form and content with KWiCFinder

KWiCFinder complements AltaVista by focusing searches to increase the relevance of webpages matched. The typical "search and select" strategy requires one to query a search engine, then retrieve and evaluate webpages one by one. KWiCFinder accelerates this operation by fetching and excerpting matching documents for the user. Even with a KWiC concordance of webpages, however, the language samples still must be considered individually and selected for usefulness.

KWiCFinder's browser-based interactive search reports allow one to evaluate large numbers of documents efficiently. The data are encoded in XML format, so results from a single search can be transformed into various "views" or formats for display in a Web browser, from "classic concordance" – one line per citation, centred on the key word or phrase – to table or paragraph layout with key words highlighted. Navigation buttons facilitate jumping from one example to the next.

In effect, KWiCFinder search reports constitute mini ad-hoc corpora which can include substantial context for further linguistic investigation. Users can add comments to relevant citations and documents, call up original or locally saved copies of webpages for further scrutiny, and select individual citations for retention or elimination from the search report. Browser-based JavaScript tools are integrated into the search report to support exhaustive exploration and simple statistical analysis of the co-text. User-enhanced search reports can be saved as stand-alone HTML pages for sharing with students or colleagues, who in turn can annotate, supplement, save and share them. By merging concordanced content and investigative software into a single HTML document that runs in a browser, KWiCFinder interactive search reports remain accessible to users of varying degrees of sophistication and achieve a significant degree of platform independence.[8]

4. Language-oriented Web search: challenges and solutions

4.1.1 Challenge I: time and effort

Each generation of computers has made us users more impatient: we have grown accustomed to accessing information instantly, and a delay of seconds can seem interminable. Tools such as KWiCFinder can download and excerpt several pages a minute, where the exact value of "several" depends on connection speed, document size and processing capability. Frequently I investigate a linguistic question or look for appropriate readings for my students by searching for and processing 100 or more webpages in 10–15 minutes. For example, to compile a sample corpus of Web documents, I downloaded 11,201 webpages in an afternoon while I was teaching through unattended simultaneous searches. Typically I run such searches while doing something else and

[8]For a discussion of features of the interactive search reports, refer to http://kwicfinder.com/KWiCFinderKWiCFeatures.html and http://kwicfinder.com/KWiCFinderReportFormats.html.

peruse the results when convenient. Unfortunately this strategy is inadequate for someone like a translator with an immediate information need, and it can be costly for a user who pays for time online by the minute. [9]

Downloading and excerpting webpages can be accelerated. In an ongoing study based on my sample Web corpus I have evaluated various "noise-reduction" techniques to improve the usefulness of documents fetched from the Web (Fletcher 2002). Document size is the simplest and most powerful predictor of usability: webpages of 3–150 KB tend to yield more connected text, while smaller or larger files have a higher proportion of non-textual overhead or noise, as well as a higher HTML-file size to text-file size ratio. Since document size can be determined before a file is fetched, one could restrict downloads to the most productive size range and achieve tremendous bandwidth savings. While this and other techniques will realize further efficiencies in search agents, even an automated search and concordancing tool like KWiCFinder remains too slow to be practical for some purposes.

Furthermore, formulating a targeted query and evaluating online documents and citations for reliability, representativeness and relevance to a specific content domain, pedagogical concern or linguistic issue can require a significant investment of time and effort. If a search addresses a question of broader interest, the resulting search report and analysis should be shared with others. While one can easily save such reports as HTML files for informal dissemination, there is no mechanism for "weblishing" them or informing interested colleagues about them. Moreover, the relevant, reliable webpages selected by a searcher are likely to lead to productive further exploration and analysis of related issues and to contain valuable links to additional resources, yet as things now stand they will be found in future searches only by coincidence. How can the value added by the (re)searcher be recovered?

4.1.2 Solution I: Web Corpus Archive (WCA)

To help searchers with an immediate information need and as a forum for sharing search results I intend to establish an online archive of Web documents which collects, disseminates and builds on users' searches. KWiCFinder will add the capability for qualified users to upload search reports with broader appeal to this Web Corpus Archive (WCA). In brief comments, users will describe the issues addressed, classify the webpages by content domain, and summarize the insights gained by analyzing the documents. Such informal weblications will enable language professionals and learners world wide to profit from an investigator's efforts. This model extends Tim Johns' concept of "kibbitzers", ad-hoc queries from the British National Corpus designed to clarify some fine point of word usage or grammar complemented by analysis and discussion of the evidence which he saves and posts online (Johns 2001).

Whenever a user uploads a search report to benefit the user community, the WCA server will download the source documents from the Web and archive them, preserving the original content from "link rot" and enabling others to verify and reanalyze the original data. Since much of a webpage's message is conveyed by elements other than raw text – images, layout, colour, sounds, interactivity – these elements should be preserved as well. Links from these pages to related content will be explored to extend the scope of content archived. Since this growing online body of webpages selected for reliability and classified by content domain will reside on a single server, it can provide fast, sophisticated searches within the WCA, yielding browser-based interactive search reports similar to those produced by KWiCFinder. Fruitless searches will be submitted to other search engines to locate additional webpages for inclusion in the Web Corpus Archive. Data on actual user searches with KWiCFinder and on my "Phrases in English" site (Fletcher 2004) would also expand the archive. Available topic recognition and text summarization software could be harnessed to classify and evaluate these automatically retrieved documents.

---

[9] I am indebted to Michael Friedbichler of the University of Innsbruck for this observation and for fruitful discussions of various issues from the user's perspective.

Clearly obtaining permission from all webpage creators to incorporate their material into an archive is unfeasible, which raises the question whether this repository would infringe on copyright. Including entire webpages without permission in a corpus distributed on CD-ROM would obviously be illegal – and unethical to boot. But providing a KWiC concordance via the Web of excerpts from webpages cached in their entirety on a "corpus server" clearly falls well within currently accepted practice. While not a legal expert, I do note that for years search engines like Google and AltaVista have included brief KWiC excerpts from documents in their search reports with impunity. Indeed, both Google and Internet Archive (a.k.a. the Wayback Machine, http://web.archive.org) serve up entire webpages and even images from their cache on demand. Both these sites' policy statements suggest an implied consent from webpage owners to cache and pass on content if the site has no standard Web exclusion protocol "robots.txt" file prohibiting this practice and if the document lacks a meta-tag specifying limitations on caching. They assert this right in daily practice and defend it when necessary in court. Internet Archive's FAQ explicitly claims that its archive does not violate copyright law, and in accordance with the Digital Millennium Copyright Act it provides a mechanism for copyright holders to request removal of their material from the site as well.[10]

Besides these familiar sites rooted in the information industry, libraries and institutes in various countries are establishing national archives of online documents to preserve them for future generations. The co-founder of one such repository (who understandably prefers anonymity) has confided in me that his group will proceed despite the unclear legality of their endeavour. Eventually legislation or litigation will clarify the status of Web archives, a recurring topic on the Internet Archive's [archivists-talk] mailing list.[11] Optimistically I assume that a Web-accessible corpus for research and education derived from online documents retrieved by a search agent in ad-hoc searches will fall within legal boundaries. Meanwhile, I intend to assert and help establish our profession's rights while scrupulously respecting any restrictions a webpage author communicates via industry-standard conventions.[12]

---

[10] In an interview (Koman 2002) Internet Archive founder Brewster Kahle brushes aside a question about copyright, insists that it is legal and implies that the Internet Archive had never had problems with any copyright holder (subsequent lawsuits nullify that implied claim). The Archive's terms of use and copyright policy also assert the legality of archiving online materials without prior permission (http://archive.org/about/terms.php [visited 8 October 2002]). Apparently such assertions are based on Title 17 Chapter 5 Section 512 of the US Digital Millenium Copyright Act (DMCA, http://www4.law.cornell.edu/uscode/17/512.html [visited 28 February 2004]), which authorizes providers of online services to cache and retransmit online content without permission from the copyright owner under specific conditions, which include publishing "takedown" procedures" for removing content when notified by the owner and leaving the original content unmodified. (Extensive discussion and documentation of these and related issues are found on the websites Chilling Effects http://www.chillingeffects.org/dmca512/ and Electronic Freedom Foundation http://www.eff.org.) Excerpting KWiC concordances from a webpage clearly constitutes modification, as does highlighting of search terms in a cached version, both services provided by Google and other search engines. Two legal experts I have consulted who requested anonymity find no *authorization* in US copyright law for these accepted practices, but case law seems to have established and reinforced their legitimacy. Obviously the legal status of these practices under US law has little bearing on the situation in other countries, whose statutes and interpretation may be more or less restrictive.

[11] In the United States, a KWiC concordance of webpages appears to fall under the fair-use provisions of copyright law as well. Crews (2000) and Hilton (2001) both argue for more liberal interpretations of this law than that found in the typical academic institution's copyright policy. I am seeking an official ruling from my institution's legal staff before establishing the WCA on servers at USNA. If our lawyers do not authorize exposing the Academy to possible risk in this gray area, I can implement the WCA on my KWiCFinder.com website. As a "company" KWiCFinder has neither income nor assets, making it an unlikely target for litigation.

[12] An approach proposed by Kilgarriff (2001), the Distributed Data Collection Initiative, would create a virtual online corpus: a classified collection of links to relevant webpages would compile subcorpora from webpages retrieved from their home sites on demand and serve them to users; as pages disappear they would be replaced by others with comparable content. This alternative avoids liability for caching implicitly

4.2.1 Challenge II: Commercial search engines

Two concerns prompt me to propose a more ambitious project as well. Firstly, the limitations imposed on queries by the most popular search engines for practical reasons reduce their usefulness for serious linguistic research. Secondly, the demands of survival in a competitive market compromise the viability and continuity of the most valuable search engines.[13]

4.2.2 Solution II: Search Engine for Applied Linguists (SEAL)

The observations in 4.2.1 point toward one conclusion: if language professionals want a search site that satisfies their needs for years to come, they will have to create and maintain it themselves. With this conviction I now outline a realistic path to this goal of a Search Engine for Applied Linguists (SEAL). While on sabbatical during the academic year 2004–05 I intend to start on this project and hope to report significant progress toward this goal at TaLC 2006.

An ideal Web search site for language learners and scholars would have to support the major written languages and character sets, and allow expansion to any additional language. The search engine would provide sophisticated querying capabilities to ensure highly relevant results, not only matching characters, but also parts of speech and even syntactic structures. Such a site would permit searches on any meaningful combination of wildcards and regular expressions, which would be optimized for the character set of the target language.[14] It also would furnish built-in language-specific "tamecards" to match morphological and orthographic variants. SEAL should not report merely how many webpages in the corpus contain a given form, but also calculate its total frequency and dispersion as well. While mainstream search engines match at the word level, ignoring the clues to linguistic and document structure contained in punctuation and HTML layout tags, our ideal site would also take such information into account. Above all, a search site for language professionals would stress quality and relevance of search results over quantity.

Real-world search sites are resource-hungry monsters. At the input end of the process, "robot" programs "crawl" or "spider" the Web, downloading webpages and adding their content to the search database. Links extracted from these documents point the way to other pages, which are spidered in turn. A "full" Web crawl involves transferring and storing many terabytes (roughly 10^{12} characters) of data. When the webpage database is completed, indexed and optimized, the search site calls on it to attend to many thousands of user queries simultaneously, with a tremendous flow of data in the other direction.

To perform their magic, major search sites boast batteries of thousands of computers, gigabytes of bandwidth, and terabytes of storage. How can we academics hope to match their capabilities? Collectively we too have thousands of computers and gigabytes of bandwidth untapped when our learning laboratories and libraries are closed. Why not employ them to crawl and index the Web for a language-oriented search engine? A central server would coordinate the tasks and accumulate the results of these armies of distributed "crawlers."

---

copyright documents, but it does not provide an instantly searchable online corpus, nor does it guarantee availability of the original data for verification and further analysis.

[13] When this was written, *AltaVista* was the only large-scale international search engine that supported wildcards and the complex queries necessary for efficient searching. Originally a technology showcase for *Digital Equipment Corporation*, it passed from one corporation to another over the years. In March 2004, the latest owner Yahoo dropped support for wildcards on the AltaVista site and apparently ceased maintaining a separate database for AltaVista. These developments reinforce my point that linguists must establish their own search engine to ensure that their needs will be met.

[14] "Regular expressions" are powerful cousins of wildcards which allow precise matching of complex patterns of characters. Unfortunately most implementations are Anglo-centric and thus ignore the fact that characters with diacritics can occur within word boundaries. Regular expression pattern-matching engines could be optimized for specific languages by matching only those characters expected to occur in a given language.

The inspiration for this distributed approach comes from a project which processes signals from outer space with a screensaver running on volunteers' desktops around the world; whenever one of the computers is idle, the program fetches chunks of data and starts crunching numbers. Researching the concept online, I discovered both a blueprint for a search engine with distributed robots spidering the Web (Melnik et al. 2001) and a Master's thesis on Herodotus, a peer-to-peer distributed Web archiving system (Burkard 2002).[15] Clearly we need not reinvent the wheel to implement SEAL, only adapt freely available open-source software to the specific requirements of our discipline.[16]

Once the basic search engine framework has been implemented and tested, the model could be extended to a further degree of "distributedness." Separate servers hosted by different universities could each concentrate on a specific language or region, or else mirror content for local users to avoid overtaxing a single server. Local linguists would provide the language-specific expertise to create tamecards for morphological and orthographic variants, optimize regular expressions for the character set, and implement part-of-speech and syntactic tagging. Due to the relatively low volume of traffic, such sites could support sophisticated processing-intensive searches which are impractical on general-purpose search engines.[17] The specialized nature and audience of a linguistic search engine *cum* archive would limit its exposure to litigation as long as the exact legal status of such services remains unclear. Indeed, since the goal of SEAL is to build a useful representative searchable sample of online documents, not to cover the Web comprehensively, some restrictions on content would be quite tolerable.

5. Alternative solutions

This section surveys existing resources comparable to those outlined above. The intention is to be descriptive, not judgemental: while a software application's usefulness for a specific purpose should be gauged by its suitability for one's goals, its success must be assessed only by how well it meets its own design objectives. The list of applications derives from variants on the question "How is your software *x* different from *y*?" Since the Web Corpus Archive and Search Engine for Applied Linguists are vapourware which may never achieve all that I intend to, I acknowledge that I am comparing an ideal concept to implemented facts.

Before a detailed discussion of the alternatives it is only fair to reveal my background, biases and intentions. Before programming a precursor of KWiCFinder in 1996, I spent 10 years designing, implementing and evaluating video-based multimedia courseware for foreign language instruction.[18] The development cycle entailed extensive direct observation of users as well as

---

[15] Links to these and other resources related to the concepts proposed here are on http://kwicfinder.com/RelatedLinks.html. In early 2003, *LookSmart*, a large commercial search engine provider, acquired *Grub*, a distributed search-engine crawling system. Now almost 21,000 volunteers have a *Grub* client screensaver which retrieves and analyzes webpages, thus helping *LookSmart* to increase the coverage and maintain the freshness of its databases. (http://looksmart.com [visited 19 June 2003]; http://grub.org [visited 19 June 2003]; http://wisenut.com, [visited 19 June 2003])

[16] Open-source software is developed cooperatively and distributed both freely and free. Specific open-source technologies proposed here are the "LAMP platform": Linux operating system, Apache web server, MySQL database, PHP and / or Perl scripting, all of which cost nothing and run competently on standard desktop PCs costing at most a few hundred dollars. Storage costs have dropped well below 50 cents a gigabyte and are set to plummet as new terabyte technologies are introduced in a few years. The expertise required to develop and maintain a search site encompasses Web protocols, database programming, and server- and client-side scripting, all skills typically available at universities.

[17] For example, due to the high processing requirements, Google—currently the most popular search engine in the world by far—does not support any wildcards, and even AltaVista restricts them severely.

[18] My experience negotiating rights to incorporate authentic video into multimedia courseware explains my hypersensitivity to copyright issues.

analysis of their errors and their evaluations of the courseware. My criteria for a good user interface were heavily influenced by Alan Cooper, who preaches that software should make it impossible for users to make errors: errors are a failure of the programmer, not the user (1995, 423–40). Usability is a primary concern in all my software development projects. For instance, studies of online search behaviour such as Körber (2000), Jansen et al. (2000) and Silverstein et al. (1999), summarized in detail in Fletcher (2001a, b), reveal that most users avoid complex queries (i.e., ones with multiple search terms joined by Boolean operators like AND, OR and NEAR), and those who do attempt them make errors up to 25% of the time, resulting in failed queries. Many features of KWiCFinder and subsequent applications address specific observed difficulties of students and other casual searchers in order to help them produce appropriate, well-formed queries.

As a teacher of Spanish and German, I sought a tool for my students and myself that could handle languages with richer morphology and greater freedom in word order than English. For example, while a typical English verb has only 4–5 variants, Spanish verbs have ten times that number of distinct forms. English sentences tend to be linear, but in German, syntactic and phraseological units are often interrupted by other constituents. In both languages webpage authors use diacritics inconsistently – Spanish-language pages may neglect acute accents, German pages may substitute *ae* for *ä* etc. and *ss* for *ß* (standard usage in Switzerland). Complex queries allowing matches with the Boolean operators NEAR / BEFORE / AFTER as well as NOT, AND and OR, tamecards for generating variant forms, and flexible character matching strategies are essential to studying these languages efficiently and effectively. None of the alternatives surveyed below offers the full range of Boolean operators and complex queries supported by KWiCFinder.

KWiCFinder was designed as a multipurpose application, to examine not just a short span of text for lexical or grammatical features, but also to assess document content and style when desired. As Stubbs (forthcoming) points out, the classical concordance line may provide too little context to infer the meaning and connotations of a word reliably. In a telling example he shows that the immediate context of many occurrences in the BNC of the phrase *horde of* appears to suggest neutral or even positive associations. The consistently negative connotations become obvious only after one examines a much larger amount of co-text. KWiCFinder's options to specify *any* length of text to excerpt and to redisplay concordances in various layouts (paragraph and table as well as concordance line) allows the flexibility to examine either the immediate or the larger context.

5.1 Web concordancer alternatives to KWiCFinder

Here "Web concordancer" is not to be understood as a Web interface to a fixed corpus like Mark Davies' (see Davies, this volume) "Corpus del español" (http://corpusdelespanol.org), the Virtual Language Centre's "Web Concordancer", (http://www.edict.com.hk/concordance/), or my "Phrases in English" site (http://pie.usna.edu), none of which features language *from* the Web; I designate the latter "online concordancers". Rather, the former are Web agents which query search engines and produce KWiC concordances of webpages matching one's search terms. The first two applications considered are commercial products, but the others were developed by and for linguists. Typically the software is installed on the user's computer (KWiCFinder, Copernic, Subject Search Spider, TextSTAT, WebKWiC), but WebCorp and WebCONC run on a Web server and are accessed via a webpage, which makes them less daunting to casual users and avoids platform compatibility issues.

For concordancing these applications follow one of three general strategies, client-side, server-side and search-engine-based processing. Client-side concordancers like KWiCFinder, Subject Search Spider and TextSTAT download webpages to the user's computer for concordancing. With a slow or expensive connection this can be a significant disadvantage, but once downloaded the texts can

be saved for subsequent examination and (re)analysis off line.[19] The server-side approach shifts the burden of fetching and concordancing webpages to the WebCorp or WebCONC server. This requires far less data transfer to the user's computer, but webpages of further interest must be fetched and saved individually by the user via the browser. Depending on the number of concurrent searches, these services can be slow or even unavailable. WebCorp does offer the option to send search results by e-mail, which prevents browser timeout and saves money for those with metered Internet access. One potential limitation of server-based processing is the unclear legality of a service which modifies webpages by excerpting them; client-side processing avoids any such risk. Search-engine-based concordancing is the fastest approach as it relies on the search engine's existing document indices; for details of the implementations, see the descriptions of Copernic and WebKWiC below.

Copernic (http://www.copernic.com) is a commercial meta-search agent which queries multiple search engines concurrently for a single word or phrase and produces a list of matching pages sorted by "relevance". While very fast, its concordances are too short and inconsistent to be useful for linguistic research; they appear to derive from the excerpts shown in search engine results. Copernic includes excellent support for a wide range of languages. The free basic version of the software evaluated constantly reminded me of the many additional features available by upgrading to one of several pay-in-advance variants. These more sophisticated products may offer the flexibility to do serious KWiC concordancing of online texts, and the high-end version (not evaluated) produces text summaries which could be useful for efficient preview and categorization of online content.

Another commercial search product, Subject Search Spider (http://www.kryltech.com), produces KWiC concordances of the search terms in a paragraph layout. All features are available in the 30-day free trial download, including full control over the number of concordances per document and the amount of context to show. SSSpider supports 34 languages, virtually all those of Europe, in addition to Arabic, Chinese, Hebrew, Japanese and Korean, and can search usenet (newsgroups) as well as the Web.[20] As with Copernic there are companion text summarization and document management suites available. One free product, SSServer, is deployed on a Web server, where it could easily be customized into an online concordancer for any of the languages supported.

WebCorp (http://www.webcorp.org.uk; Morley, Renouf and Kehoe 2003) from the University of Liverpool's Research and Development Unit for English Studies has regularly added new features since its launch in 2000. While it offers but a single field for inputting search terms, its support for wildcards and "patterns" (similar to KWiCFinder's tamecards) gives it flexibility in matching variant forms, and queries can be submitted to half a dozen different search engines to improve their yield. Up to 50 words of preceding and following context are shown, and options allow displaying any number of concordances per document (up to 200 webpages maximum are analyzed). WebCorp's concordances give access to additional data analysis (e.g., type / token count, lists of word forms), and other tools are available on the site. Online newspapers can be searched by domain (e.g., UK broadsheet, UK tabloid, US), and searches can be limited to a specific Open Directory content domain. With the numerous choices WebCorp offers, its failure to provide a document language option seems inexplicable, since almost every search engine supports it. The user interface would benefit from client-side checking for meaningful, well-formed queries before submission to WebCorp; mistakes in a query can lead to long waits with no results and no explanation. Zoni (2003) describes WebCorp in greater detail and compares it with KWiCFinder.

---

[19] KWiCFinder provides the option to save the Web document files automatically in original HTML and / or text format for later analysis by a full-featured concordancer like WordSmith or MonoConc.

[20] SSSpider's heuristics for determining the language of the source text are not entirely reliable: a search for pages in Afrikaans returned many Dutch pages; after switching to a search term that does not exist in Dutch, I got pages in French and Romanian as well as Afrikaans.

Matthias Hüning's WebCONC (http://www.niederlandistik.fu-berlin.de/cgi-bin/web-conc.cgi), another server-based Web concordancer, performs searches on Google and generates KWiC concordances of the search phrase in paragraph layout. One can also copy and paste text for concordancing onto the search page. Options are minimal: target language, amount of context (maximum of 50 *characters* before / after the node!), and number of webpages to process (50 maximum, in practice fewer if some pages in the search results are inaccessible or do not match exactly). There is no provision for wildcards (not supported by Google) or pattern matching. Matches are literal, and all occurrences of a search string are highlighted in the results, even as a substring of a longer word. A punctuation mark after word form is matched too, which can be useful, for example to find clause final verb forms in German. The server can be slow and may even time out without producing any concordances. WebCONC could be far more useful if it offered more options for search and output format. Of greater potential interest is the author's TextSTAT package (http://www.niederlandistik.fu-berlin.de/textstat/software-en.html), which can download and concordance both webpages and usenet postings. Programmed in Python, it runs on any standard platform (Windows, Macintosh, Unix / Linux). Hüning's user license permits modification and redistribution of the software code, making TextSTAT an instructive example and valuable point-of-departure for a customized Web concordancer.

WebKWiC (http://kwicfinder.com/WebKWiC/; Fletcher 2001 a, b) is a browser-based application that exploits a feature of Google's search results: click on the "cache" link to see a version of a webpage from Google's archives with the search terms highlighted. WebKWiC queries Google, parses the search results, fetches a page from Google's cache, encodes the highlighted search terms to permit navigation from one instance of the search terms to the next, and displays the page in a new browser window. This "parasitic" approach with JavaScript and DHTML builds on core functionalities of Internet Explorer, works on multiple platforms, and supports any language known to Google. It could be extended to produce and display KWiC excerpts from webpages, or to download and save them in HTML, text or concordance format. A small set of webpages and scripts (70KB installed), WebKWiC takes full advantage of all Google's search options.

5.2 Alternative to the Web Corpus Archive

The Internet Archive (http://web.archive.org) "Wayback Machine" preserves many (but by no means all) webpages back to 1996. Archived sites are represented by a selection of their pages and graphics in snapshots taken every few months. For example, a visit to the Archive reminded me that KWiCFinder was not publicly downloadable until November 1999, and it helped me reconstruct the introduction and evolution of WebCorp. The archive is not searchable by text, only by URL. The ability to step back in time, for example, to retrieve a webpage cited in this paper which has since disappeared from the Web, is complemented by comparison of various versions of the same webpage, with the differences highlighted. In contrast to the Internet Archive, the WCA proposed here will not aim to preserve the state of the entire Web, only to ensure immediate text-searchable access to pages which support either a user-uploaded "kibbitzoid" search analysis or documents indexed in its Search Engine for Applied Linguists.

5.3 Alternatives to the Search Engine for Applied Linguists

GlossaNet (http://glossa.fltr.ucl.ac.be) analyzes text from over 100 newspapers in eleven languages, providing both more and less than a linguistic search engine as I conceive it. Originally a monitoring tool to track emerging lexical developments (Fairon and Courtois 2000), GlossaNet now offers both "instant search" of the current day's newspapers with results in a webpage and "subscription search" (after free registration), which re-queries each daily crop of newspapers and e-mails the results at regular intervals. Concordance lines display 40 characters to the left and right of the node. Clicking on the node displays the original newspaper article with the search terms highlighted, but this feature may be unavailable: an error message warns that most articles

are accessible only on the day of publication. Queries can be formulated as any combination of word form, lemma, "regular expression" (less than the name suggests), or word class and morphology, or else as a Unitex "finite state graph" (not documented on the site; manual in French and Portuguese at http://www-igm.univ-mlv.fr/~unitex/). GlossaNet has its limitations: it is restricted to a single genre, newspaper texts, and to the rather small pool (in comparison to the Web) of one day's newspapers; searches cannot be replicated on another day, and results may not be verifiable in the context of the original article; syntactic analysis and lemmatization can be faulty; search results do not show the grammatical annotation, so the users cannot learn to tailor their queries to the idiosyncrasies of the analysis engine; documentation is minimal. Clearly it has strengths as well compared to KWiCFinder or WebCorp: the ability to search by syntactic or morphological category can eliminate large numbers of irrelevant hits; "instant search" delivers results almost immediately; "subscription search" permits monitoring of linguistic developments in manageable increments; newspaper texts are generally reliable, authoritative linguistic sources.

The Linguist's Search Engine (LSE, http://lse.umiacs.umd.edu:8080) arrived on the scene in January 2004 as a tool for theoretical linguists to test their intuitions by "treating the Web as a searchable linguistically annotated corpus" (Resnick and Elkiss 2004). At its launch LSE had a collection of about 3 million English sentences, a number bound to increase rapidly. The source of these Web documents is the Internet Archive, which ensures their continued availability. New users will likely start with the powerful "Query by Example" feature: enter a sentence or fragment to match, then click "Parse" to generate both a tree and a bracketed representation of the example sentence. LSE uses a non-controversial Penn Treebank-style syntactic constituency annotation readily accessible to most linguists. Queries can be refined in either the graphical tree or the text bracketed representation. For example, I entered "He is not to be trusted", which yielded this parse in bracketed notation:
(S1 (S (NP (PRP He)) (VP(AUX is) (S (RB not) (VP (TO to) (VP (AUX be) (VP (VBN trusted)))))))).
After being made more general in the tree editor, the bracketed query
(S1(S NP (VP(AUX be)(S(RB not)(VP(TO to)(VP(AUX be)(VP VBN)))))))
matched 76 sentences with comparable constructions such as *"Any statements made concerning the utility of the Program are not to be construed as express or implied warranties."* and *"In clearness it is not to be compared to it."*

LSE's concordances can be displayed or downloaded in CSV format for importation into a database or spreadsheet, and the original webpages can be retrieved from the Internet Archive for examination. While such linguistic search of a precompiled Web corpus via an intuitive user interface is impressive, the LSE really advances Web searching by exploiting this functionality to locate examples matching lexical and syntactic criteria with the AltaVista search engine. The user submits a query to AltaVista and LSE fetches the corresponding webpages, parses them, and filters out the ones that fail to meet the user's structural criteria. Retrieval and analysis are surprisingly rapid. Queries, their outputs, and the original webpages can be saved in personal collections for later analysis and refinement. The tools can also analyze corpora uploaded from the user's computer.

Despite the LSE's impressive power and usability, it does not fulfil all the needs the SEAL intends to address. Above all it supports only English, and there are no plans to add other languages except possibly in parallel corpora searchable via the annotation of the corresponding English passages (Resnick, personal communication), while SEAL will start with the major European languages, establish a transferable model for branching out into other language families. LSE is aimed at theoretical linguists seeking to test syntactic hypotheses who are sufficiently motivated to master a powerful but complex system. In contrast, SEAL's target audience is more practically oriented, including language professionals such as instructors, investigators and developers of teaching materials, translators, lexicographers, literary scholars, and advanced foreign language learners as well as linguists. Many in these groups could be overwhelmed by a resource that

requires too much linguistic or technical sophistication at the outset. SEAL will offer tools to leverage users' familiarity with popular search engines and nurture them along the path from word and phrase search to queries that match specific content domains, phrases structures and sentence patterns as well. As an incrementally implemented companion to the Web Corpus Archive, it will benefit both from analysis of search behaviors and use patterns and from direct user feedback. After comparing future plans, Resnick and I have determined that LSE and SEAL will complement rather than compete with each other.

6. Web search resources in language teaching and learning

Suggestions for language teachers and learners to use these tools are surveyed here. Specific examples of instructor-developed learning activities focussing on the levels of word, phrase and grammar are based on my experience teaching beginning and intermediate German and Spanish. Open-ended learner-directed techniques to develop critical searching skills and to encourage writing by example are also described. While some of these tasks could be performed without the specialized software described here, they make the process more effective and familiarize the students with valuable research tools and techniques applicable to other disciplines as well.

Since 1996 the Grammar Safari site (http://www.iei.uiuc.edu/web.pages/grammarsafari.html) has been a popular resource on the Web, linked to and expanded on by over 2000 other sites. It offers tutorials and a set of assignments for learners of English to hunt for and analyze grammatical and rhetorical structures in online documents. The technique entails querying a search engine, retrieving webpages individually, and finding the desired forms on the page. One of the Web concordancers surveyed above could easily automate the mechanics of such activities, leaving more time for analysis and discussion of the examples. Familiarizing learners with an efficient approach to a beneficial but tedious task will encourage them to apply it even when not directed to do so.

Grammatical and lexical exploration can also be based on instructor-prepared mini-corpora. KWiCFinder allows search results to be saved as webpages with self-contained interactive concordance tools which can be used profitably with students. For example, to contrast the German passive auxiliary *wurde* with the subjunctive auxiliary *würde,* I assign small groups of students (2-3 per computer) to explore, then describe the grammatical context (e.g., they co-occur with past participle and infinitive respectively) to the class. As instructor I clarify the meaning and use of the structures by translating representative examples. These few minutes spent on "grammar discovery" prepare the students to understand and retain the textbook explanation better.

Recently an in-class KWiCFinder search demonstrated to my students how actual usage can differ from textbook prescription. In a geographical survey of the German-speaking countries I explained that the usual adjective for "Swiss" in attributive position is the indeclinable *Schweizer;* a student pointed out that our textbook listed only *schweizerisch.* A pair of KWiCFinder searches rapidly clarified the situation: while forms of the latter typically modified the names of organizations and government institutions, the former was obviously both far more frequent and more general in use. Students can be assigned similar ad-hoc discovery activities in response to recurrent errors or to supplement the textbook. For example, it is instructive for a learner studying French prepositions to discover that *merci* ***à*** */* ***pour*** parallel English "thanks **to** / **for**", while *merci* ***de*** *+ infinitive* corresponds to English "thanks **for"** *+ -ing.* A search of the BNC illustrates the advantage of a bottomless corpus like the Web: this English construction occurs only 53 times in this huge corpus, and could well be lacking entirely in a smaller one. Ideally, after assigned tasks such as this, learners will develop the habit of formulating and verifying usage by example rather than resorting to Babelfish or another online translation engine.

Frand (2000) summarizes what he calls the "mindset" of Information-Age students. Their behavior with an unfamiliar website or software package typically exhibits more action than reflection; learning by trial and error replaces systematic preparation and exploration ("Nintendo over logic"). To encourage development of "premeditated" searching habits, I assign students a written pre-search exercise before they undertake open-ended Web-based research for a report or essay. They jot down variants of key words and phrases likely to occur on webpages in contexts of interest for their topic as well as additional terms that can help restrict search results to relevant webpages.[21] This written exercise forces thought to precede action and allows the group to brainstorm about additional possible search terms and variants. Then they search for and evaluate a number of webpages in writing with a checklist based on Barton 2004. Finally, they re-search the sites deemed most useful in order to find additional appropriate content. Without these paper-and-pencil exercises, students tend to choose from the first few hits for whatever search term occurs to them. A concordancing search agent greatly accelerates evaluating webpages for content, reliability, and linguistic level.

One venerable stylistic technique I attempt to pass on to my students is *imitatio* (not plagiarism!), the study and emulation of exemplary (or at least native speaker) texts in creative work. In major languages the Web is a generous source of texts on almost any topic. After locating appropriate webpages, advanced learners can immerse themselves in the style and language of the content domain they are dealing with before preparing compositions or presentations. This concept parallels translation techniques outlined by Zanettin (2001) based on ad-hoc corpora from the Web. It is a powerful life-long foreign-language communication strategy which builds knowledge as well as linguistic skills.

When the WCA and SEAL and comparable resources become a reality, they will further accelerate the tasks surveyed here. Response time from querying from a single archive will be far faster than fetching and excerpting documents from around the Web. Searching a large body of selected documents by content domain and / or grammatical structure will yield a higher percentage of useful hits than the current query by word form approach. User-submitted kibbitzers will supply ready illustration and explanation for linguistic questions and problems (e.g., the *wurde / würde* and *Schweizer / schweizerisch* examples above).[22] Finally, the linguistic annotation provided by SEAL will help motivated students gain greater insight into grammar.

Admittedly, most of the techniques discussed here are feasible with static corpora as well. By the same token, most applications of corpus techniques to language learning (surveyed in Lamy and Mortensen 2000) could be adapted to Web concordancing instead. The size and comprehensive coverage of the Web are powerful arguments for this approach, as is the availability of free tools with a consistent, adaptable user interface for exploring everything from linguistic form to document content. If we can acquaint our students with responsible online research techniques and instil in them a healthy dose of skepticism toward their preferred information source, we will have accomplished far more than teaching them a language.[23]

---

[21] KWiCFinder's inclusion and exclusion criteria are terms which help narrow but are not concordanced in the search results. For example, in a search for *TaLC,* words like "corpus", "corpora", "language", "linguistic*s*" are good discriminators of relevant texts, while "powder, talcum" are likely to appear on irrelevant webpages.

[22] Perhaps Philip King's term "kibbitzoids", premiered at TaLC 5 in Bertinoro (2002), is more appropriate, as these are not strictly speaking what Tim Johns means by kibbitzers.

[23] As Frand (2000:16) puts it, "Unfortunately, many of our students do believe that everything they need to know is on the Web and that it's all free."

7. Caffè e grappa oppure limoncello

In this paper we have considered a wide range of challenges and solutions to exploiting the Web as a (source of) linguistic corpus. Such dense, heavy fare leaves us much to digest. Let's linger over *caffè* and *grappa* or *limoncello* to discuss these ideas - after all, this is not just a declaration of intent, but an invitation to a dialogue.

These proposals outline an incremental approach to implementing the solutions which will yield useful results at every milestone along the way - searchers with an immediate information need should not have to delay gratification as a programmer must. The Web Corpus Archive proposed here will give direct search results, if not the first time, then at least when a query is submitted on subsequent occasions. Posted KWiCFinder search report kibbitzers can exemplify techniques for finding the forms or information one requires, much as successful recipes from a pot-luck supper continue to enrich the table of those who adopt them.

Building on the infrastructure of this archive, the Search Engine for Applied Linguists sketched here will afford rapid targeted access to an ever-expanding subset of the Web. In the process, all three information-gathering strategies will be served: hunters will profit from a precision search tool, grazers will be able to locate rich pastures of related documents, and browsers will enjoy increased likelihood of serendipitous finds. As other linguists join in the proposed cooperative effort, the search engine's scope can be extended well beyond European languages. Initially, outside funding may be required to establish the infrastructure, but ultimately this plan will be sustainable with resources from the participating institutions.

With time, the incomparable freshness, abundant variety and comprehensive coverage added by this Web corpus-*cum*-search engine will make it an indispensable complement to the more reliable canned corpora for a "pick'n'mix" approach. Linguists and language learners alike will benefit from examples which clarify grammatical, lexical or cultural points. Foreign language instructors and translators will find a concentrated store of useful texts for instructional materials and translation by example. New software tools will integrate the Web and the desktop into a powerful exploratory environment. The steps outlined here will lead toward fulfilling the Web's promise as a linguistic and cultural resource.

References

Aston, G. 2002. "The learner as corpus designer". In *Teaching and Learning by Doing Corpus Analysis: Proceedings of the Fourth International Conference on Teaching and Language Corpora, Graz 19-24 July, 2000* [Series Language and Computers, Vol. 42], B. Kettemann and G. Marko (eds), 9-25. Amsterdam: Rodopi.

Banko, M. and Brill, E. 2001. "Scaling to very very large corpora for natural language disambiguation". ACL-01. Online at http://research.microsoft.com/~brill/Pubs/ACL2001.pdf [verified 2.3.2004]

Barton, J. 2004. "Evaluating web pages: techniques to apply & questions to ask". Online: http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/Evaluate.html [visited 1.3.2004]. Evaluation form online: http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/EvalForm.pdf [visited 1.3.2004].

Brekke, M. 2000. "From the BNC toward the cybercorpus: a quantum leap into chaos?" In *Corpora Galore: Analyses and Techniques in Describing English: Papers from the Nineteenth International Conference on English Language Research on Computerised Corpora (ICAME 1998),* Kirk, J. M. (ed.), 227-247. Amsterdam and Atlanta: Rodopi.

Burkard, T. 2002. "Herodotus: a peer-to-peer web archival system". Cambridge, MA, Massachusetts Institute of Technology Master's Thesis. Online: http://www.pdos.lcs.mit.edu/papers/chord:tburkard-meng.pdf [visited 8.10.2002].

Burnard, L. and McEnery, T. (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.

Cooper, A. 1995. *About Face: The Essentials of User Interface Design.* Foster City, CA: IDG Books.

Crews, K.D. 2000. "Fair use: overview and meaning for higher education". Online: http://www.iupui.edu/~copyinfo/highered2000.html [visited 8.10.2002].

De Schryver, G. M., 2002. "Web for / as corpus: a perspective for the African languages". *Nordic Journal of African Studies* 11(2): 266-282. Online at http://tshwanedje.com/publications/webtocorpus.pdf [verified 26.2.2004].

Fairon, C. and Courtois, B. 2000. "Les corpus dynamiques et GlossaNet: Extension de la couverture lexicale des dictionnaires électroniques anglais". JADT 2000 : 5es Journées Internationales d'Analyse Statistique des Données Textuelles. Online: http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2000/pdf/52/52.pdf [visited 18.2.2003].

Fletcher, W. H. 2001a. "Re-searching the web for language professionals". CALICO, University of Central Florida, Orlando, FL, 15-17 March 2001. PowerPoint online: http://www.kwicfinder.com/Calico2001.pps [visited 2.3.2004].

Fletcher, W. H. 2001b. "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. Online: http://kwicfinder.com/FletcherCLLT2001.pdf [visited 8.10.2002].

Fletcher, W. H. 2002. "Making the web more useful as a source for linguistic corpora". American Association for Applied Corpus Linguistics Symposium, Indianapolis, IN, 1-3 November 2002. Online: http://kwicfinder.com/FletcherAAACL2002.pdf [visited 25.8.2003].

Fletcher, W. H. 2004. "Phrases in English". Online database for the study of English words and phrases at http://pie.usna.edu [visited 26.2.2004].

Frand, J. 2000. "The Information-Age mindset: Changes in students and implications for higher education". *EDUCAUSE Review* 35 (5): 14-24. Online: http://www.educause.edu/pub/er/erm00/articles005/erm0051.pdf [visited 29.2.2004].

Ghani, R., Jones, R. and Mladenic, D. 2001. "Using the web to create minority language corpora". 10th International Conference on Information and Knowledge Management (CIKM-2001). Online at http://www.cs.cmu.edu/~TextLearning/corpusbuilder/papers/cikm2001.pdf [visited 7.7.2003].

Grefenstette, G. 1999. "The World Wide Web as a resource for example-based machine translation tasks". Online at http://www.xrce.xerox.com/research/mltt/publications/Documents/P49030/content/gg_aslib.pdf [visited 12.10.2001]

Hilton, J. 2001. "Copyright assumptions and challenges". *EDUCAUSE Review* 36 (6): 48-55. Online: http://www.educause.edu/ir/library/pdf/erm0163.pdf [visited 8.10.2002].

Hofland, K. 2002. "Et Web-basert aviskorpus". Online: http://www.hit.uib.no/aviskorpus/ [visited 8.10.2002].

Jansen, B. J., Spink, A. and 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.

Johns, T. F. 2001. "Modifying the paradigm". Third North American Symposium on Corpus Linguistics and Language Teaching, Boston, MA, 23-25 March 2001.

Kilgarriff, A. 2001. "Web as corpus". In *Proceedings of the Corpus Linguistics 2001 conference, UCREL Technical Papers: 13,* P. Rayson, A. Wilson, T. McEnery, A. Hardie and S. Khoja, (eds), 342-344. Lancaster: Lancaster University. Online: http://www.itri.bton.ac.uk/~Adam.Kilgarriff/PAPERS/corpling.txt [visited 8.10.2002].

Kilgarriff, A. and Grefenstette, G. 2003. "Introduction to the special issue on the web as corpus". *Computational Linguistics* 29 (3): 333-47. Online: http://www-mitpress.mit.edu/journals/pdf/coli_29_3_333_0.pdf [visited 7.1.2004].

Körber, S. 2000. "Suchmuster erfahrener und unerfahrener Suchmaschinennutzer im deutschsprachigen World Wide Web: ein Experiment". Unpublished master's thesis, Westfälische Wilhelms-Universität Münster, Germany. Online: http://kommunix.uni-muenster.de/IfK/examen/koerber/suchmuster.pdf [visited 9.1.2004].

Koman, R. 2002. "How the Wayback Machine works". Online: http://www.xml.com/pub/a/ws/2002/01/18/brewster.html [visited 8.10.2002].

Lamy, M. N. and Mortensen, H. J. K. 2000. "ICT4LT Module 2.4. Using concordance programs in the modern foreign languages classroom". *Information and Communications Technology for Language Teachers.* Online: http://www.ict4lt.org/en/en_mod2-4.htm [visited 1.3.2004].

Melnik, S., Raghavan, S., Yang, B. and García-Molina, H. 2001. "Building a distributed full-text index for the web". WWW10, 2-5 May 2001, Hong Kong. Online: http://www10.org/cdrom/papers/275/index.html [visited 8.10.2002].

Morley, B., Renouf, A. and Kehoe, A. 2003. "Linguistic research with XML/RDF-aware WebCorp tool". http://www2003.org/cdrom/papers/poster/p005/p5-morley.html [visited 19.2.2004].

Pearson, J. 2000. "Surfing the Internet: teaching students to choose their texts wisely". In Burnard and McEnery, 235-239.

Resnik, P. and Elkiss, A. 2004. "The Linguist's Search Engine: getting started guide". http://lse.umiacs.umd.edu:8080/lse_guide.html [visited 23.1.2004].

Scannell, K. P. 2004. "Corpus building for minority languages". Online at http://borel.slu.edu/crubadan/ [visited 19.3.2004].

Silverstein, C., Henzinger, M., Marais, H. and Moricz, M. 1999. "Analysis of a very large web search engine query log". *SIGIR Forum,* 33(3). Online at http://www.acm.org/sigir/forum/F99/Silverstein.pdf [verified 26.2.2004].

Smarr, J. 2002. "GoogleLing: the web as a linguistic corpus". Online at http://www.stanford.edu/class/cs276a/projects/reports/jsmarr-grow.pdf [visited 7.7.2003].

Stubbs, M. forthcoming. "Inferring meaning: text, technology and questions of induction". In *Aspects of Automatic Text Analysis,*. R. Köhler and A. Mehler (eds). Heidelberg: Physica-Verlag.

Varantola, K. 2003. "Translators and disposable corpora". In *Corpora in Translator Education,* F. Zanettin, S. Bernardini and D. Stewart (eds), 55-70. Manchester: St Jerome.

Volk, M. 2002. "Using the web as corpus for linguistic research". In *Tähendusepüüdja: Catcher of the Meaning. A Festschrift for Professor Haldur Õim* [Publications of the Department of General Linguistics 3], Pajusalu, R. and Hennoste, T. (eds.). Tartu: University of Tartu. Online at http://www.ifi.unizh.ch/cl/volk/papers/Oim_Festschrift_2002.pdf [visited 7.7.2003].

Zanettin, F. 2001. "DIY corpora: the WWW and the translator". In *Training the Language Services Provider for the New Millennium*, B. Maia, H. Haller and M. Urlrych (eds), 239-248. Porto: Facultade de Letras, Universidade do Porto. Online: http://www.federicozanettin.net/DIYcorpora.htm [visited 16.12.2003].

Zoni, E. 2003. "e-MINING - Software per concordanze online". Online: http://applicata.clifo.unibo.it/risorse_online/e-mining/e-Mining_concordanze_online.htm [visited 15.1.2004].