

Corpus Analysis of the World Wide Web

William H. Fletcher
United States Naval Academy
fletcher at usna dot edu

to appear in Chapelle, Carol A, (Ed.). (2011). *Encyclopedia of Applied Linguistics*. Wiley-Blackwell.
<http://www.encyclopediaofappliedlinguistics.com/>

Introduction

The World Wide Web has become a primary meeting place for information and recreation, for communication and commerce for a quarter of the world's population. Millions of Web authors have created billions of webpages, unknowingly providing texts to be mined for their linguistic and cultural content. The Web has evolved into the resource of first resort for lexicographers and linguists, for translators, teachers and other language professionals. As a source of machine-readable texts for corpus linguists and researchers in complementary fields like Natural Language Processing (NLP), Information Retrieval and Text Mining, the Web offers extraordinary accessibility, quantity, variety and cost-effectiveness. Investigators in these disciplines have developed scores of tools and products from Web content for both researchers and end users and authored hundreds of scholarly papers on their projects.

This article reviews the rewards and limitations of either acquiring Web content and processing it into a static corpus or else accessing it directly as a dynamic corpus, a distinction captured in *Web for / as corpus* (De Schryver, 2002); here *Web corpus* serves as a cover term for both approaches. In the process it surveys typical applications of such data to both academic analysis and real-world situations, discusses tools and techniques available to motivated language professionals and learners, and outlines future directions for Web corpus development.

Advantages and Limitations of Web Corpus

The Web corpus owes its popularity to its tremendous size; broad linguistic, geographic and social range; up-to-dateness; multimodality; and wide availability at minimal cost. We shall consider each advantage briefly before discussing the limitations of the Web corpus.

Size

The Web universe is constantly expanding, so its size is unknowable. In 2008 Google noted that it had identified (but not actually indexed) over a trillion (10^{12}) distinct URLs (Web addresses), and that several billion (10^9) new webpages appear daily (Alpert & Hajaj, 2008). Estimates suggest that Google indexes about 40 billion webpages; the search engine (SE) Cuil.com's home page claims to index 127 billion, albeit with much repeated content. Size is about more than impressive statistics: Halevy, Norvig and Pereira illustrate "the unreasonable effectiveness of

[huge sets of] data” derived from the Web when applied to NLP problems like automatic speech recognition (ASR) and machine translation, arguing that “...invariably, simple models and a lot of data trump more elaborate models based on less data” (2009, 8-9), an insight reflected in the steadily improving quality of their employer Google’s online translation and speech transcription tools. Size also matters because many words and phrases are relatively infrequent, so even large traditional corpora like the British National Corpus (BNC, 100 million words) offer few examples if any of familiar constructions to be found in abundance online.

Range

Once overwhelmingly Anglophone and Eurocentric, the Web has become truly World Wide, encompassing most of the earth and its written languages; English-language users and content now have minority status online (Top Ten Languages—Internet World Statistics, 2009). The availability of webtext has stimulated the development of corpora, lexica and NLP tools for under-resourced languages across the globe, including Africa (<http://AfLaT.org>) and Southeast Asia (<http://SEAlang.net>). Internationally-oriented sites provide multilingual material which can aid translators and language learners. Online social networks have redefined our notion of community, bringing formerly private interaction into public view. In one example, such open access to informal discourse permitted Thelwall (2008) to compare swearing on the social site MySpace by males and females from the U.K. and the U.S. for teenagers and middle-aged adults; users unintentionally supplied both linguistic and demographic data.

Up-to-dateness

Major SEs closely track unfolding events and incorporate content from online communities with minimal time lag, enabling real-time study of current topics and emerging usage. Indeed, the Web and associated technologies have been both the catalyst for much linguistic creativity and the main vehicle for its dissemination. The Web is the most effective resource to research and document linguistic reflections of contemporary culture. In contrast, any static corpus is cut off at the moment of its compilation. For example, since the enormous Google 1T corpus, based on a trillion words of online English, was completed shortly before the appearance of the social site Twitter, it lacks any trace of the hundreds of playful neologisms coined by “tweeple”.

Multimodality: sound, sight and text

Video and audio represent a sizable segment of online content, but search and linguistic analysis such as part-of-speech (PoS) tagging and concordancing require machine-readable written text. Some commercial news broadcasts and interviews are available in transcript form (Hoffmann, 2007), affording access to scripted and unscripted language in both spoken and written form. Sites which provide searchable (and downloadable) transcripts of audio content via ASR have come and gone, but new technologies are emerging to provide comparable services for online video: Exalead’s Voxalead, Google Audio Indexing, YouTube ASR captioning and translation, and, most impressively, IBM TALES, which supports search with English queries for video content in

other languages and provides reasonably accurate original and machine-translated text. Such sites enable Web corpora which combine video or audio with written text.

Availability

In the developed world, internet access is fast and inexpensive, and the growing demand in developing countries continues to improve service and reduce costs there as well. With an inexpensive PC and a home broadband connection one can compile and process a multimillion-word corpus in minutes. Besides Web-only content, such a corpus can contain material available in other forms, such as articles from newspapers and magazines, entire books and even movie subtitles. Thanks to the Web, acquisition of computerized text in almost any modern written language has become a straightforward task.

Limitations of Web corpus

Despite the obvious benefits, Web content has drawbacks and limitations as linguistic data. Unless one restricts oneself to preselected sites, the source and linguistic properties of a webpage remain unknown. For whom and what purpose is the text intended? What geographic and demographic origin and target audience does it represent? Was it written carefully or carelessly by a native speaker, or is it an unreliable translation by man or machine? Is the document authoritative – accurate in content and representative in linguistic form? Typically a Web corpus cannot answer such questions, key elements in the design of traditional corpora. Other caveats and limitations are discussed below.

Web as / for corpus (WaC / WfC)

Web as corpus

The most widespread approach to accessing the Web as a corpus is through commercial SEs like Google, Yahoo! or Bing. For some purposes a quick glance at snippets of search terms in context on a search engine results page (SERP) suffices. For further insight users can retrieve webpages from the SE's cache to display all occurrences of the word forms matched highlighted in context, a rudimentary but rapid sort of concordance.

In another scenario a user queries a SE for alternate forms, phrases or spellings, then compares the number of hits reported; Googlefight.com animates such comparisons. Not surprisingly, some linguists have based serious discussions on the comparison of “ghits” (Google hits), an unreliable practice for several reasons (Kilgarriff, 2006). Hit counts are very different from corpus frequencies: they reflect the number of matching webpages, not actual occurrences, and a single page may be counted for alternate search terms. Many webpages duplicate each other's content, inflating the counts. Indeed, “most widespread” does not necessarily mean “preferred”: “ungrammatical” German *da werden Sie geholfen!* ‘You (*formal*) will be helped there!’ gained popularity as an advertising slogan, so it is far more frequent online than “correct” *da wird Ihnen geholfen!* In contrast, the “grammatically correct” analog with informal *dir* wins this Googlefight

by a five-to-one ratio. Similarly, reinterpretations of common English phrases like *straight-laced*, *hone in on*, *font of wisdom / knowledge* rival or exceed the accepted spellings in frequency online.

Hits tally more than exact matches, further skewing page counts. To identify potential hits in major languages which do not match a query exactly, SEs typically apply *stemming*, i.e. reduction to a base form, erasing distinctions among e.g. verb forms and even verbs and related nouns, so a search for *governed* also matches pages with *governing* or *government*. The dominant Russian SE Yandex.com performs similar simplification of morphologically far more complex languages like Russian and Kazakh, one factor in its success in the region. Some webpages are included in hit counts not for their own text but for the content of pages linking to them. SEs ignore punctuation, so some reported hits for a phrase might be spread across a phrase or sentence boundary. Diacritics are treated inconsistently or simply disregarded. Most importantly, reported hit counts only roughly approximate the actual number of matching documents and can vary greatly even within a few minutes. Particularly misleading is the “union of sets” problem: hit count algorithms typically underestimate results for an OR query, reporting fewer matches for alternate search terms than for one of the terms alone. (For details of such issues see Uyar, 2009 a & b; Thelwall, 2008b; Fletcher, 2007a; Duclaye, Collin & Pétrier, 2006.)

As gatekeepers of the Web, the major SEs also have limitations for users who want to explore the language behind the hit counts. They are useful for linguistic research by coincidence, not design, and their services are subject to change or cancellation without notice. Since they exist to steer impatient searchers toward any relevant webpages, major SEs limit results to the top 1000 matching URLs as ranked by proprietary schemes, which tend to favor media and similar popular sites, and often tailor result sets to give prominence to the most popular matches and to reflect the user’s geographic location and past search activity. Such second-guessing of the searcher’s intent can skew results and mask the full range of possible matches.

Despite such limitations, the three major SEs do afford effective access to a significant subset of the Web: only they have the massive infrastructure required to discover, index, store and retrieve data from billions of webpages efficiently. By delivering results which favor text-rich pages and authoritative sites and reduce the number of duplicate and questionable documents, they raise the quality of SERP content. Several free tools facilitate WaC research via these SEs. For example, WebCorp (<http://www.webcorp.org.uk>; Renouf, Kehoe & Banjeree, 2007) provides an online service to query SEs, then generate concordances, word lists and document metrics from matching webpages. KWicFinder (<http://kwicfinder.com>; Fletcher, 2007a) runs on a PC to produce concordances from webpages; it can save local copies of pages retrieved for further analysis or inclusion in an offline corpus. While these tools are optimized for European languages, some SE-intermediary sites such as the SouthEast Asian Language Web Corpus (<http://SEALang.net>) focus on other regions. In contrast, WebAsCorpus.org (Fletcher, 2007b) concordances webpages in any language and character set. To enable rapid creation of ad-hoc corpora, this site compresses matching webpages into a single zipfile for download and offline analysis.

All three of these applications provide rapid answers to questions of language usage with examples highlighted in context, but the total number of matching documents is limited by the SEs. To build larger Web corpora, the BootCaT software package (<http://bootcat.sslmit.unibo.it>; Baroni & Bernardini, 2004) starts from a user-provided list of seed terms and iteratively queries

a SE, retrieves documents, extracts related terms, then requeries the SE with these new terms and repeats the process. Another free suite of tools, Jaguar, extracts specialized corpora from the Web and analyzes various lexical features statistically (Nazar, Vivaldi & Cabré, 2008). Its document-clustering techniques reduce extraneous content by filtering webpages based on their similarity to user-designated documents. Since Jaguar provides data processing and storage, it places fewer demands on the user's technical expertise and infrastructure.

Tools like WebCorp and WebAsCorpus.org are regularly used by translators, writers and language learners to verify usage and refine and enrich their writing. Language teachers enlist them to find authentic examples and appropriate texts for their students. Of course, such applications require the user to identify concerns, initiate a search and evaluate the result. A new generation of SE-based tools for foreign language professionals and learners aims to support the writing process transparently. For example, Microsoft's Web-based ESL Assistant detects a number of common learner errors (e.g. articles, prepositions, word order) and suggests improvements based on both large-scale language models derived from Web data and real-time SE queries (Gamon, Leacock, Brockett, Dolan, Gao, Belenko & Klementiev, 2009). It highlights potential problems and furnishes examples of alternative formulations along with confidence scores derived from SE hit counts.

Other important SEs could also be tapped by applications like those described above. They provide alternative gateways to the Web with different features, search results and focuses. For example, Exalead.com still supports in-word wildcards and proximity operators (* and NEAR), features dropped by AltaVista in 2004. Baidu.com, the leading SE in China, indexes content from Asian countries not accessible via other SEs. There are hundreds of niche SEs for specific languages, regions or content areas which can be harnessed for WaC.

Web for corpus

"Crawling" or "spidering" the Web to compile a do-it-yourself corpus offers a powerful but challenging alternative to SE-mediated access. Starting from a list of "seed" URLs, crawler software "harvests" webpages, saves copies locally, extracts links, downloads new webpages linked to, then repeats the process until user-defined criteria are met. Crawlers can be restricted to desired domains (e.g. UK, RU) or websites (e.g. a specific newspaper, government or social site), and limits can be set on the thoroughness with which each site is harvested.

After downloading, webpages must be processed for importation into corpus analysis tools and databases. This entails converting the various document formats found on the Web (HTML, PDF, DOC etc.) into clean plain text, a very challenging process, especially for languages with multiple encoding schemes for non-Roman scripts (e.g. Khmer, Burmese, Japanese), and detecting the language to filter out extraneous material. To reduce noise and improve representativeness one may also remove "boilerplate" or "template" elements (content recurring on each page, navigation links etc.), nearly identical or highly repetitive documents, and webpages not consisting primarily of coherent text (Fletcher, 2004; Baroni and Kilgarriff, 2006; Scannell, 2007). To enhance usefulness for linguistic research, the texts may also be tagged with

grammatical information and / or normalized, e.g. converted to lower case or corrected for typographical and spelling errors.

Many software applications are available to crawl the Web and process webpages into usable text, but such solutions typically require programming expertise to customize and coordinate the various processes. A rare exception is multiplatform GrosMoteur (<http://grosmoteur.elizia.net>), which produces concordances either by spidering the Web from a list of seed URLs or by querying Yahoo! directly. For the more technically proficient, Web Archive's Heritrix crawler and complementary applications offer greater flexibility and power.

Some free academic sites support Web corpus creation and annotation by harvesting online content on demand and processing it according to the user's specifications. Glossa.net calls itself a "watch engine": a user registers a query for specific terms or structures in a variety of language and receives concordances enhanced with morphological, syntactical and semantic information as Glossa.net encounters them. Alternatively one can take advantage of numerous free application programming interfaces (APIs) to Web services to perform sophisticated tasks through simple scripts adapted from sample code. For example, AlchemyAPI retrieves and processes webpages, supporting language detection, text extraction (content minus boilerplate) and document categorization by topic. uClassify offers predefined document classifiers (e.g. detection of language, author gender) as well as a customizable module which could be "trained" to classify texts by genre or topic.

The following paragraphs survey representative examples of Web corpora acquired through crawling. The Crúbadán Project (Scannell, 2007) is compiling corpora for 431 under-resourced languages to enable translation and adaptation of open-source end-user applications (word processors, spellcheckers) for those languages. The Leeds Collection of Internet Corpora (Sharoff, 2006) provides a search interface to large (up to 100 million words) linguistically annotated corpora in a dozen European and Asian languages. The WaCky Project has compiled tagged corpora of one-two billion words each for several European languages available for free download to academic researchers (Baroni, Bernardini, Ferraresi & Zanchetta, 2009). Much larger tagged corpora with target sizes of 10-20 billion words have been developed, including WebCorp's Linguists Search Engine (Renouf, 2009) and BiWeC (Pomikálek, Rychlý & Kilgarriff, 2009).

Lexicographers have been avid consumers of Web data. For example, the Web-derived two-billion-word Oxford English Corpus was built to support dictionary development and to monitor the evolution of English. The University of Leipzig's Wortschatz-Portal (<http://wortschatz.uni-leipzig.de/>) has compiled online-searchable lexica and downloadable corpora for 57 languages. For German this portal offers a free API to obtain detailed lexical, morphological, syntactic and semantic information as well as examples of usage for any given wordform. Specialized aids for translators, writers and language learners would be easy to develop with this service. Web data can also support bilingual lexicography: Ferraresi, Bernardini, Picci and Baroni (2008) show that typical collocations found in large independently-compiled corpora of English and French can assist lexicographers in revising a bilingual dictionary.

A Web crawl can also target only a specific site or set of sites. For example, Gutenberg.org, Wikibooks.org and other eBook sites are unrestricted sources of literary, academic and general-

interest texts. Many media companies have online archives of articles and transcripts which can be crawled; the most impressive compilations from such sources are Mark Davies' Time Corpus and Corpus of Contemporary American English (<http://corpus.byu.edu/>). For languages with few NLP tools available, crawling sites with high-quality text can provide the data for producing them; for example, to develop a lexicon and PoS-tagger for Bengali, Ekbal and Bandyopadhyay harvested a Web corpus from the archives of a single newspaper (2008).

To develop translation aids, websites with multilingual content can be mined for parallel corpora. WeBiText.com has compiled a "Translation Memory" of searchable parallel texts in 30 languages, mainly from EU and other government sites (Désilets, Farley, Patenaude & Stojanovic, 2008); SERPs show translation equivalents in context. Linguee.de targets a single language pair, German-English, and draws on corpora from a wider range of bilingual sites spanning most content areas; SERPs highlight equivalent terms and indicate level of confidence in their accuracy based on quality of the sources and user ratings. Of course, as Jiménez-Crespo (2009) demonstrates, human-translated sites also are subject to strong influence from the source language. Consequently, corpora with comparable content drawn from similar sites in different language areas may be preferable to parallel corpora in which one text is inevitably translated. For a study of language and culture in Italy and the UK as reflected in the phraseology of agritourism, Manca (2008) compiled monolingual comparable corpora from sites in both countries; the contrasts her analysis reveals would not have emerged as clearly from translated texts.

Such investigator-compiled Web corpora remain orders of magnitude smaller than the SEs' databases. To stimulate research, Google has released data from two sets of Web corpora, one of over a trillion tokens of English, the other of around 100 billion tokens each in 10 other European languages. One prototype project, Linggle, enhances Google's data with PoS tags and supports search in this vast dataset by wildcard and / or wild-PoS to identify recurring collocations (Chang, 2008); for example, **/a beach* returns collocations of *adjective beach* by actual frequency (not document frequency) in Google's database. Wu, Witten and Franken (2010) describe a much more ambitious ongoing project to combine filtered Google data with text examples from the BNC in a variety of English learning activities. Another application finds an approach to detecting real-word spelling errors (e.g. mistaken *there* for *their*) based on Google's datasets more effective than other spellcheckers (Islam & Inkpen, 2009).

Copyright

Copyright issues remain a gray area in compiling and distributing Web corpora. In principle anything found on the Web is copyright and therefore subject to control by its author. Free access to Web content does not imply the right to download, retain, reprocess and redistribute it. A small percentage of webpages do explicitly permit reuse and redistribution under Copyleft or Creative Commons license; Yahoo! and Google both support searching only such pages. Obtaining permission from all copyright holders to reuse texts from a large webcrawl borders on the impossible. The interpretation of international copyright law varies widely by country, ranging from unrestricted use for research purposes to "no use is fair use"; Hemming and Lassi (n. d.) discuss the issues without resolving them. The labor-intensive nature of postprocessing data from a webcrawl and the need of investigators to share data for verification and replication

of their results by other researchers argue for redistributing Web corpora even if it entails some legal risk.

Trends in Web Corpus Development

The Web as corpus community continues to grow in size and productivity. The Association for Computational Linguistics Special Interest Group on Web as Corpus (SIGWAC) organizes annual workshops which attract both corpus linguists and NLP researchers. Web data are increasingly prominent in papers in other related fields as well. To the enormous benefit of corpus linguists, the SE industry's research agenda overlaps significantly with their own. Investigators from the major SEs regularly publish papers with insights from their research on Web-scale data, and the industry also supports development of open-source software which scales better to corpora with billions of tokens than do traditional database systems. This software includes Lucene, a powerful full-text SE, and Hadoop, a framework to distribute processing of data-intensive applications over many PCs, as well as more specialized tools based on them like Mahout, a distributed machine-learning library for document clustering (identifying similar documents) and classification (assigning documents to predetermined categories). The ability not only to compile but also to process, store, manage and uncover meaningful patterns in huge datasets benefits investigators and end users alike.

References

- Alpert, J. & Hajaj, N. (2008, July 25). We knew the web was big... [Web log post]. Retrieved from <http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html>
- Baroni, M. & Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, 1313–1316.
- Baroni, M. & Kilgarriff, A. (2006). Large linguistically-processed web corpora for multiple languages. *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics*, 87–90.
- Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation Journal*, 43(2), 209–226.
- Chang, J. S. (2008). Linggle: a web-scale language reference search engine. Unpublished manuscript.
- De Schryver, G.-M. (2002). Web for / as corpus: a perspective for the African languages. *Nordic Journal of African Studies*, 11, 266–282.

- Désilets, A., Farley, B., Patenaude, G., & Stojanovic, M. (2008). WeBiText: Building large heterogeneous translation memories from parallel web content. *Proceedings of the ASLIB International Conference Translating and the Computer 30, London: ASLIB*.
- Duclaye, F., Collin, O., & Pétrier, E. (2006). Fouille du Web pour la collecte de données linguistiques: avantages et inconvénients d'un corpus hors-normes. *Actes de l'atelier « Fouille du Web » des 6èmes journées francophones « Extraction et Gestion des Connaissances »*, Lille, 53–64.
- Ekbal A., & Bandyopadhyay S. (2008). Web-based Bengali news corpus for lexicon development and POS tagging. *Polibits* (37).
- Fairon, C., Macé, K. & Naets, H. (2008). GlossaNet 2: A linguistic search engine for RSS-based corpora. In S. Evert, A. Kilgarriff, A. & S. Sharoff (Eds.). *Proceedings of the 4th web as corpus workshop (WAC-4)*, pp. 34–39.
- Ferraresi, A., Bernardini, S., Picci, G. & Baroni, M. (2008, August). Comparable web corpora for bilingual lexicography: a pilot study of English / French collocation extraction and translation. Symposium *Using Corpora in Contrastive and Translation Studies*, Zhejiang University, China.
- Fletcher, W. H. (2004). Making the web more useful as a source for linguistic corpora. In U. Connor & T. Upton (Eds.), *Applied corpus linguistics: a multidimensional perspective* (pp. 191–205). Amsterdam: Rodopi.
- Fletcher, W. H. (2007a). Concordancing the web: Promise and problems, tools and techniques. In M. Hundt, N. Nesselhauf & C. Biewer (Eds.), *Corpus linguistics and the web* (pp. 25–45). Amsterdam: Rodopi.
- Fletcher, W. H. (2007b). Implementing a BNC-compare-able web corpus. In C. Fairon, H. Naets, A. Kilgarriff & G.-M. De Schryver (Eds.), *Building and exploring web corpora* (pp. 43–56). Louvain-la-Neuve: Cahiers du Cental.
- Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24 (2), 8–12.
- Hoffmann, S. (2007). From webpage to mega-corpus: the CNN transcripts. In M. Hundt, N. Nesselhauf & C. Biewer (Eds.), *Corpus linguistics and the web* (pp. 69–85). Amsterdam: Rodopi.
- Islam, A. & Inkpen, D. (2009). Real-word spelling correction using Google Web 1T 3-grams. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 1241–1249.

- Hemming, C. & Lassi, M. (n. d.) Copyright and the Web as Corpus. Presentation, Stockholm University. Retrieved from <http://hemming.se/gslt/copyrightHemmingLassi.pdf>
- Gamon, M., Leacock, C., Brockett, C., Dolan, W. B., Gao, J., Belenko, D. & Klementiev, A. (2008). Using statistical techniques and web search to correct ESL errors. *CALICO Journal*, 26(3), 491–511.
- Jiménez-Crespo, M. A. (2009). Conventions in localisation: a corpus study of original vs. translated web texts. *The Journal of Specialised Translation*, 12, 79–102.
- Kilgarriff, A. (2006). Googleology is bad science. *Computational Linguistics* 33(1), 147–151.
- Leturia I., Gurrutxaga A., Alegria I. & Ezeiza A. (2007). CorpEus, a 'web as corpus' tool designed for the agglutinative nature of Basque. In C. Fairon, H. Naets, A. Kilgarriff & G.-M. De Schryver (Eds.), *Building and exploring web corpora* (pp. 69–81). Louvain-la-Neuve: Cahiers du Cental.
- Manca, E. (2008). From phraseology to culture: Qualifying adjectives in the language of tourism. *International Journal of Corpus Linguistics*, 13(3), 368–385.
- Nazar, R., Vivaldi, J. & Cabré, M. T. (2008). A suite to compile and analyze an LSP corpus. *Proceedings of LREC 2008*, 1164–1169.
- Pomikálek, J., Rychlý, P. & Kilgarriff, A. (2009). Scaling to billion-plus word corpora. *Advances in Computational Linguistics*, 41, 3–13.
- Renouf, A., Kehoe, A. & Banerjee, J. (2007). WebCorp: An integrated system for web text search. In M. Hundt, N. Nesselhauf & C. Biewer (Eds.), *Corpus linguistics and the web* (pp. 47–67). Amsterdam: Rodopi.
- Renouf, A. (2009). Corpus linguistics beyond Google: the WebCorp Linguist's Search Engine. *Digital Studies / Le champ numérique*, 1(1), (not paginated).
- Scannell, K. P. (2007). The Crúbadán Project: Corpus building for under-resourced languages. In C. Fairon, H. Naets, A. Kilgarriff & G.-M. De Schryver (Eds.), *Building and exploring web corpora* (pp. 5–15). Louvain-la-Neuve: Cahiers du Cental.
- Sharoff, S. (2006). Open-source corpora: Using the net to fish for linguistic data. *International Journal of Corpus Linguistics*, 11(4), 435–462.
- Thelwall, M. (2008a). Fk yea I swear: Cursing and gender in a corpus of MySpace pages. *Corpora*, 3(1), 83–107.
- Thelwall, M. (2008b). Quantitative comparisons of search engine results. *Journal of the American Society for Information Science and Technology*, 59(11), 1702–1710.

Top Ten Languages – Internet World Statistics. (2009, December 31). Retrieved from <http://www.internetworldstats.com/stats7.htm>

Uyar, A. (2009a). Investigation of the accuracy of search engine hit counts. *Journal of Information Science*, 35(4), 469–480.

Uyar, A. (2009b). Google stemming mechanisms. *Journal of Information Science*, 35(5), 499–514.

Wu, S., Witten, I. H. & Franken, M. (2010). Utilizing lexical data from a web-derived corpus to expand productive collocation knowledge. *ReCALL*, 22(1), 83–102.

[Suggested Readings]

Baroni, M. & Bernardini, S. (Eds.). (2006). *WaCky! working papers on the web as corpus*. Bologna: GEDIT.

Bergh, G. & Zanchetta, E. (2008). Web linguistics. In A. Lüdeling & M. Kytö (Eds.) *Corpus Linguistics. An international handbook*. Berlin: Mouton de Gruyter (Vol. 3), 309–327.

Gatto, M. (2009). *From body to web: an introduction to the web as corpus*. Bari: Editori Laterza.

Kilgarriff, A. & Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29(3), 333–348.

Lindquist, H. (2009). Corpus linguistics in cyberspace. In *Corpus linguistics and the description of English* (chap. 10, pp. 187-206). Edinburgh: Edinburgh University Press.

Online Resources

Resources for “Corpus Analysis of the World Wide Web”, <http://webascorpus.org/EAL/>