BOOK REVIEWS


Web as Corpus is a much welcomed book because it is the first unified account of the role of the web in the now well-established discipline of corpus linguistics.

Until this monograph, issues variously pertaining to the relation between language corpora and the World Wide Web had been raised by a number of researchers that were too busy developing specific aspects of that relation to provide the whole picture. Web as Corpus indeed gives readers the whole picture: it offers a very thorough and engaging explanation of the emergent philosophy and revolutionary role played by the web in corpus linguistic study; it contains cutting-edge research without indulging in too technical details; and it enables the student novice to actually make use of the web for linguistic purposes. Not only is the argument supported throughout by compelling examples, but also each chapter is complemented with a “Study questions and activities” section, which is extremely instructive for the novice and the expert researcher alike. I have tried most of the activities myself to gain first-hand experience and they all work very well in helping readers through this new field of enquiry. What follows is a more detailed review of the contents of each chapter.

The topic of the book (web as corpus, or, is the internet a suitable linguistic corpus?) is introduced in an original way by delving right into the still open debate on corpus linguistics as theory or method. The author builds on that to deal with “a number of methods that look at the web as their main resource to implement the corpus linguistic approach” (p. 7). She then overviews the fundamental tenets of corpus linguistics, including corpus authenticity, representativeness, balance and sampling, size and composition, and outlines the analytical tools that are commonly used to generate and analyse corpus data, i.e. word lists, keyword lists, and concordances. These are illustrated through a number of sample searches mainly from the BNC (British National Corpus). The overview is functional to the discussion in the subsequent chapters.

It is, however, in Chapter 2 where the epistemological positioning of corpus linguistics (CL) on one hand and the web as corpus on the other is explored and the difference made clear. The defining features of the “traditional” corpus as a “body of text” introduced in Chapter 1 are revisited in light of the challenges posed by the web “as a spontaneous self-generating collection of texts” (p. 35, p. 41, but see
also p. 73 and p. 211 “the web as a huge self-generating collection of authentic text in machine-readable format”). In this chapter the author engages with the critique that, although the web shares some of the defining features of traditional corpora as collections of searchable authentic texts in machine readable form, it is volatile and ephemeral, occasionally anonymously or multiply authored, thus posing questions of accuracy and authoritativeness. Even more crucially, it has certainly not been put together for a linguistic purpose and therefore seems to lack the features of sampling and representativeness that typify a carefully compiled corpus. How can one talk about design criteria, asks Gatto, with “a database as anarchic and chaotic as the World Wide Web” (p. 36)?

The anarchy of the web is thus analysed in detail in terms of its contents, the language they are in, their medium, the domains covered and the genres and text types into which they can be classified (pp. 49–63). The very dynamism of these contents, the scope and variety of the texts, the immense size of the data available, which, in the same way as with a monitor corpus, tends to correct any skew and achieve representativeness, lead the author to conclude that what from a more traditional corpus perspective may seem like limitations are instead the strengths of an alternative web as corpus model. In this model, as she convincingly argues, the web reflects the social dimension and natural dynamism of language which is always in a state of flux, never in a state of being, thus becoming almost a metaphor for “discourse” (pp. 66–68).

The analysis of the web’s contents contained in this chapter also prompts a useful explanation of two important aspects of Information Retrieval, namely ‘precision’ and ‘recall’ (also, ‘relevance’ and ‘reliability’). As done with other fundamental concepts so far, these too are discussed in light of the workings of the web and related to one of the topics covered by the three central chapters, specifically dealing with tools for the creation of corpora from the web (the web for corpus approach).

I have dwelt on Chapter 2 because I find it especially important in that, in introducing the challenges posed by the web as corpus, it offers an opportunity to rethink issues that even the corpus-savvy reader might take for granted. The chapter anticipates the conclusion of the book that CL and studies on the web as corpus are bound to coexist and interact for a long time to come (see also the Conclusions on p. 212). Such fruitful interaction is expressed metaphorically by analogy with the “force field” in physics, a connection generated through attraction, which is created between the two approaches (p. 41).

Chapters 3 to 5 map the evolution of the web as corpus approach by explaining the use of the web for linguistic analyses via ordinary search engines (Chapter 3) and the web as corpus vs. web for corpus distinction (Chapter 4 and Chapter 5). Each chapter corresponds to a stage in the linguist’s analytical path when dealing
with web data, it is therefore rich in exemplification. The first step consists in using commercial search engines (e.g. Google) to interrogate the web in order to retrieve linguistic information. This is said to have been widely employed by computational linguists to address Natural Language Processing tasks like Machine Translation and Word Sense Disambiguation tasks, as well as corpus linguists working with language learners, for whom the web is an invaluable source of attested language use (pp. 73–74). The chapter addresses the inherent limitations which were already pointed out in Chapter 2, notably lack of control on the part of the linguist over such a “huge, anarchic and unstable database” (p. 74), hybridity in terms of medium, occasional uncertainty about authorship, and the resulting difficulty in assessing data statistically, and provides operational solutions accompanied by a large number of illustrations. One of the major limitations of commercial search engines is that search criteria do not match linguistically oriented analyses. The retrieval algorithms are designed in a way that is meant to enhance effectiveness of the results they yield for commercial purposes, inevitably reducing linguistic relevance. Gatto explains that this happens as a consequence of normalisation of spelling variants, lemmatisation and stemming, and the inclusion of synonyms in the results (p. 75), but also as a consequence of the evolving nature of the queries the web is navigated for, which are context-driven and meant to target user’s behaviour (indeed something we frustratingly experience in our everyday surfing of the web!).

The core of Chapter 3 is therefore devoted to introducing the basic functioning of web search engines as well as the advanced search options that are of use to the linguist and language professional. Search engine basics, i.e. crawling via a web crawler, indexing with an indexing programme, searching via a query processor and ranking of results (pp. 77–83), will particularly interest the reader who is not aware of how general web search works. However, it is the more advanced query that the user who is interested in finding linguistic information needs to be able to formulate. Here the author makes the important point that in placing even more emphasis on the query, its formulation and the evaluation of results, web as corpus differs significantly from traditional corpus research.

A number of examples of how to query the web for linguistic information are given through a detailed illustration of the advanced search options in Google. These include Boolean operators and the wild card (p. 84 ff.), restrictions as to national web domains as well as academic or genre specific domains, and date range specification “to turn the web into a sort of diachronic corpus” (p. 87). The query so formulated allows the user to compensate for the lack of representativeness and unreliability of the web and is therefore shown to be absolutely essential to the exploratory process of linguistic analysis. The sample queries include collocations and phraseological expressions, the latter being fruitfully searched for by
combining phrase search with the wild card (see, for example, the search for the pattern to * a better future on p. 92), which are tested on domain-specific subsections of the web (see the example of if, as has been argued restricted to the site: .ac. uk only, p. 94).

Gatto underlines the obvious advantage over traditional corpora of searching for phraseology on the web. It is well-known to linguists that when searching for phraseological expressions or collocations, one needs more data. Therefore, compared to any other corpus, however big, the web is likely to yield many examples of long phrases that are relatively sparse in other corpora. At the same time, the user is warned by the author that results should be handled with care, given that even less frequent, but attested collocations might appear among the search results. This is why the author observes that there is also a lot of noise on the web of which one should be aware and “aim not so much at validating the initial hypothesis, but rather at invalidating it” (p. 89). Among the examples of the fruitful interaction of corpus linguistic analysis and Google services are the ngrams facility (p. 94) and Mark Davies’ (2014) Google Books interface (p. 95), which are also clearly illustrated by Gatto.

The chapter ends with searches for translation candidates, another way to profitably exploit the web affordances as a corpus (pp. 96–101). The idea here is that the vast amount of data from different languages that is available on the web can be used by the linguist and translation professional to check possible translation candidates, just by looking at their relative frequency and by restricting the search to domain-specific subparts of the web. This might concern alternative syntactic formulations (e.g. is site of onset a better translation than onset site for the Italian expression “sede di insorgenza” in medical research? See pp. 97–98) or preferred collocates (e.g. the Italian “paesaggi aspri” is best translated into English as rugged landscapes, the most quantitatively relevant and qualitatively appropriate bigram in original English texts/webpages, see pp. 99–101). This final example is chosen to show how the process of gradual query refinement is essential to obtaining maximum relevance and reliability.

The last paragraph in this chapter should not be overlooked as it makes the parallel between web searches and traditional corpus searches explicit: the search for a single word is to the web as the compilation of a finite size corpus is to CL research (see “it is as if, albeit only for a few moments, our virtually endless corpus, the web, complies with finite size, one of the fundamental criteria of corpus design”, p. 101). Similarly, the search for two or more words is a way to create co-text and add some kind of a topic or domain restriction to the use of a given word. Along the same lines, other restrictions to the search in terms of URL and web domains have their correlate in CL searches for geographical, genre and register variation.
A further stage in the linguist’s approach to web data is expounded in Chapter 4, covering the customised concordancing tools which have been developed by linguists to create concordances directly from the web. Unlike commercial search engines, these tools generate concordances in the linguist’s typical KWIC format (thus allowing resorting of results and various other computations) and can incorporate linguistic queries and filter options. The example considered in detail is that of WebCorp Live and the WebCorp Project developed by Andrew Kehoe and Antoinette Renouf (see Renouf et al. 2007) at Birmingham City University. A number of case studies that make use of this tool are reported, usefully testing collocational profiles, neologisms (e.g. the blend flexicurity) and lexical creativity (pp. 106–118). Because WebCorp Live is a free tool, it is definitely a resource that readers will be able to use in their research. Another web concordancing tool, WebAsCorpus, associated with the name of William Fletcher (Fletcher 2004), is also described despite the fact that it is no longer functioning, for it displays interesting features.

Reflecting the above-mentioned analytical movement from web as corpus or web as corpus surrogate (an idea that the author draws from Baroni & Bernardini (2006), whose work is referred to throughout the book) to web for corpus, Chapters 4 and 5 introduce and illustrate more tools, which have been developed to create ad hoc offline corpora drawing from data on the web and to interrogate them. In particular, the functionalities of WebCorpLSE (an acronym for Linguist’s Search Engine developed by the same team at Birmingham City University, see Kehoe & Gee 2007) and BootCaT (Baroni & Bernardini 2004) are illustrated both through a number of significant case studies in the literature and with material original to the author’s own investigations, which stand out as being especially efficacious didactically. Through the WebCorpLSE site, the user has free access to three very diverse web corpora including a large synchronic corpus of English, a diachronic one, and a corpus of blog texts that is especially useful to those interested in the study of CMC, Computer Mediated Communication (see the example of kinda sorta on pp. 131–133). The example of the phrase credit crunch is given to illustrate language variation over time using the diachronic web corpus (see pp. 129–130).

The examples chosen to illustrate the functioning of BootCaT in Chapter 5 are taken from the field of Terminology and Translation and show how to create specialised corpora and term-lists from the web, thus fully embodying the web for corpus approach. Gatto explains that the corpus creation process is started up through a set of words or phrases that the researcher considers likely to occur in a given domain (the so-called ‘seeds’, following BootCaT terminology). Starting from the inputted seeds, the tool automatically downloads the top webpages returned by a search engine for each query and, after post-processing them, produces the ad hoc corpus available for analysis offline. This in turn can trigger the
identification of further seeds that can be iteratively used to obtain new webpages and expand the corpus. In this way, the procedure is recursive and can be repeated until the corpus reaches the desired size. Also, it allows user control over the kinds of filters that are associated with each query, the number of queries and seeds used per query, etc. Gatto notices that if on the one hand it is easy, and certainly quick, to compile specialised corpora, on the other, the BootCaT bootstrapping procedure places the entire burden of the corpus construction on the researcher who will have to carefully choose the seeds and discard the irrelevant URLs returned by the system in a cyclical fashion.

This point brings us back to the importance of query formulation with respect to precision and recall, another fundamental lesson I learn from this book. The statement on p. 102 that “it is only through a process of progressive query refinement towards greater complexity that linguists can contribute to improve the quality of the results” applies here as well, where the corpus user coincides with the corpus builder (something which does not necessarily happen with traditional corpora). This is a theme that runs through Gatto’s treatment of the subject at hand, and one that is especially important when training users not just of the web as corpus, but of any corpus. In a way, what the author is saying is that the web can become a corpus and can be taken as such if you know how to deal with it by paying attention to search criteria and the heuristics typical of CL. The focus is thus shifted from the web as a reservoir of text to how you deal with this reservoir, that is the investigative operations and procedures.

Following the same procedure, BootCaT can be used to compile comparable corpora in more than one language, so that equivalences and differences between specialised terms can be explored (see the example of the Italian translation “potenza installata” as the best equivalent for the English installed capacity and, working in the opposite direction, from Italian into English, wind farm as the most appropriate translation for “impianto eolico”, pp. 156–158). The protocol described is of use to translation professionals, trainers and trainees alike, in that it helps them document the corpus creation process, formulate queries, evaluate results and process web documents for the purposes of specialised translation.

The next chapter (Chapter 6) is concerned with reviewing web corpora for languages other than English and some of the projects associated with them, notably the TenTen collection, which refers to corpora from a variety of languages of ten billion words each (for developments of these corpora, Gatto refers readers to the Sketch Engine website). Furthermore, the chapter describes the functioning of another tool, whose latest applications were also developed within the web as corpus paradigm, namely Sketch Engine (Kilgarriff et al. 2004), to reproduce the standards of traditional corpus linguistic research with huge amounts of web data. In particular, one of the functions of the Sketch Engine exemplifies how the web
as corpus can be used not just for language analysis but also for understanding something of a culture and society at large (as announced by the captivating title of the chapter, “Sketches of Language and Culture from Large Web Corpora”). The author here explores the word *culture*, exploiting to the full the properties of the web as corpus (pp. 183–202).

Finally, Chapter 7 looks at the future by pointing to possible developments of the web as corpus brought about by the technology of the so-called second generation web. The case considered is that of Wikis as repositories of user-generated content and particularly Wikipedia, whose multilingual composition has been exploited for the compilation of comparable and parallel corpora in some very recent experiments. The other technological innovations that are likely to affect web as corpus include cloud computing and cooperative programming, especially as far as production, distribution and access of the resources are concerned. Indeed, an outcome of Gatto’s in-depth exploration of the role of the web in language studies, and one I find especially valuable for the student and the researcher alike, is that while you learn about web as corpus, you also understand how the internet as such, with its latest generation affordances, works. Given the pervasive nature of the web in this digital era we live in, to provide readers with a broader understanding of the issues at stake seems as important as to lead them through the specialisms of the discipline.

In conclusion, this book shows the complementarity between CL and web as corpus, and the mutual enrichment that derives from their interaction. It encompasses a wealth of resources and illustrations, while at the same time it provides a full theoretical interpretation of the changes brought about by the web in linguistic research. For all the above reasons, Gatto’s *Web as Corpus* is already among the set readings of my Corpora and Language Variation course reading list for the academic year just started.

References


Reviewed by Maria Freddi, University of Pavia, Italy