DE GRUYTER OPEN

***Review in ICAME Journal*, Volume 39, 2015, DOI: 10.1515/icame-2015-0008**

Maristella Gatto. *Web as corpus: Theory and practice.* London and New York: Bloomsbury. 2014. 232 pp. ISBN 978-1-4411-5098-1 (hardback); ISBN 978-1-4411-6112-3 (paperback); ISBN 978-1-4411-3413-4 (E-book). Reviewed by **Mirka Honkanen**, University of Freiburg.

This monograph presents a highly accessible, concise, and vivid discussion of the main issues of using the web as a corpus, ranging from the theoretical implications it has for corpus linguistics as a field to the various tools and methods that have been appropriated and developed by web corpus linguists over the past fifteen years. The successful combination of theory and practice by the author is complemented by a brief historical survey of the development of the web corpus approach. The book is meant as an introduction to the subject for students in any language-related programmes, and as a resource for "teachers or researchers in the humanities" more generally; accordingly, the focus is on how to use existing tools rather than how to create new ones (p. 1). To be sure, *Web as corpus* is suitable for readers with limited technical skills and little previous knowledge of corpus linguistics, but more experienced corpus linguists are likely to find a few jewels in there as well. The book comprises seven chapters, of which the first two are introductory and the last one is very short. Each chapter is followed by study questions and activities, which mainly do not go beyond the text itself, and a useful list of suggestions for further reading.

The first chapter explains the author's position that corpus linguistics is an "approach" with certain theoretical implications about language and with its own methods (p. 6). The bulk of the chapter deals with the basics of corpus linguistics: key concepts are explained together with the different steps of corpus creation, some basic analyses typically conducted on corpora, and certain statistical formulae. Space is also given to a discussion of accepted standards in corpus compilation, particularly the use of authentic language, representativeness, balance (in terms of topics, genres, publishing dates, and mediums), and large enough database size. The initial chapter attends to the needs of readers with no previous contact with corpora, but is largely superfluous to experienced corpus linguists.

Chapter 2 turns attention to the web from a corpus linguistic perspective. Gatto is setting the stage for the remainder of the book by promising to deal with three out of Baroni and Bernardini's (2006: 10–14) four conceptualizations of "the web as/for corpus":

(1) "the Web as a corpus surrogate": using the web like a corpus, exploring it through a search engine – either a commercial one or one designed more specifically for linguists' needs;

(2) "the Web as a corpus shop": collecting "disposable corpora" from the web either manually or semi-automatically;

(3) "the mega-Corpus mini-Web": a new way of combining web-derived and corpus-like features, explained thoroughly only later in the book. (p. 37)

The fourth meaning, "the Web as corpus proper", which has Internet language use as the object of study, is excluded from the book completely on the basis that "it would require – if appropriately pursued – a book-length study in itself" (p. 2). This is perhaps true, but this choice weakens Gatto's claim to have produced a comprehensive treatment (p. 1). It is interesting that this particular area has been left out, when Gatto herself later admits, quoting Leech (2007: 145), that the web "can in no way be considered a representative sample of language use in general" (p. 44).

Using the web as a corpus, in certain ways, stretches our conception of what a corpus is: "the traditional notion of a linguistic corpus as a *body* of texts rests on some correlate issues, such as finite size, balance, part-whole relationship and permanence; on the other hand, the very idea of a *web* of texts brings about notions of non-finiteness, flexibility, de-centring/re-centring, and provisionality" (p. 35). Gatto discusses the theoretical consequences of the web as a corpus approach, revisiting the standards introduced in Chapter 1 one by one. The author finds pros and cons for using the web as a corpus for each of the criteria discussed. Firstly, web texts are clearly authentic communication, but their sloppiness in terms of spelling and grammar at times gives the material a feeling of un-authoritativeness. Secondly, although web corpora are routinely criticized for not being representative (a next-to-impossible goal for any corpus, it should be added), Gatto argues that online activities do represent a significant portion of human interaction, but without being "constructed by a human mind" like traditional corpora are (p. 44). Thirdly, the immense size of the web is one of its clearest advantages from a corpus-perspective, but ist vastness and uncontrollable growth pose notable methodological and theoretical challenges. The inclu-

sion of the size estimates of the web from 10–15 years back that Gatto gives at this point is of questionable value. Finally, the last section on "composition" does not have a direct counterpart in the previous chapter, but it discusses various anarchic aspects of the web as corpus, such as relation to traditional mediums of spoken and written language, multilingualism, topic and genre classification, and copyright issues.

After discussing these challenges inherited from traditional corpus linguistics, Gatto tackles a number of altogether "new issues" (p. 65 onwards) that recent technological advances and the web as corpus have brought along, corresponding partially to the changes predicted by Wynne (2002: 1204) for language in the 21st century more generally; these are the dynamism of content, the non-reproducibility of results, and the difficulty of obtaining as much relevant data as possible without collecting too much irrelevant data (the common problem of recall vs. precision).

Chapters 3 and 4, launching the more practically oriented part of the monograph, explore Baroni and Bernardini's concept of the "Web as a corpus surrogate" (2006), that is, how to use online search engines to imitate the functions of concordancing software. Chapter 3 focuses on traditional search engines, such as Google Search. After again going through some benefits and drawbacks of the web as a linguistic resource, Gatto presents us with various well-known problems with using commercial search engines for research purposes. Their functioning principles are explained, after which a considerable portion of the chapter is dedicated to practical advice and examples of how to manipulate the advanced search functions of search engines in order to avoid certain biases inherent in these systems and to obtain more relevant results. According to Gatto, the web can be usefully explored via search engines; for example, to test collocations, translation candidates, or the existence of certain linguistic forms. As the examples chosen demonstrate, these methods are perhaps most useful to translators and language learners.

The fourth chapter continues where the third left off, introducing some tools developed to serve linguists' needs better than search engines with commercial interests do. Some of these resources mainly rearrange the search results into a form that is more useful for linguists. An example is WebCorp *Live*, which returns the results in the Key-Word-in-Context format familiar from offline concordance programmes. An interesting feature offered by WebCorp *Live* are collocational profiles of words, presented as tables of collocates and their positions in relation to the node word. For some reason, WebCONC and WebAsCorpus concordancing tools are also introduced, although they had already become unavailable when the book was written. The examples and case studies come

again from a language learning context, from collocations, neologisms, and phraseology. The many figures and tables are well-explained throughout the book and serve their purpose of illustrating the various methods introduced excellently, despite a few minor errata in the captions ('Science' instead of 'Chemistry' on p. 59, 'landscapes' instead of 'landscape design' on p. 82, 'emission' instead of Italian 'emissioni' on p. 157, a differing frequency value in one of the figures and in the text on p. 177, and a few other formatting or typographic issues).

Towards the end of Chapter 4, Gatto introduces a "linguistic search engine" project which developed the WebCorp Linguist's Search Engine (WebCorp*LSE*) (p. 122). This represents a move towards the "mega-Corpus mini-Web" approach, as the programme downloads corpora from the web and then updates them with new data regularly. The project gathered three corpora: the Synchronic English Web Corpus, the Diachronic English Web Corpus, and the Birmingham Blog Corpus, which can be accessed online through a linguist-friendly interface and are suitable for "more sophisticate [sic] research questions" (p. 134). Gatto sees the WebCorp*LSE* as a best-of-both-worlds resource, stating that "like a corpus, this is a body of text of known size and composition, which is available for offline processing and analysis; like the web, it is very large and constantly updated" (p. 123).

Chapter 5 is also remarkable and potentially useful for more advanced corpus linguists. It explores the "Web as a corpus shop" application by Baroni and Bernardini (2006), explaining how small specialized *ad hoc* corpora can be collected semi-automatically using the freely downloadable BootCaT software, or its online version WebBootCaT. BootCaT only needs a list of key-words or word pairs ("seeds") which occur frequently in texts from the chosen special field. It then downloads a certain number of pages for each query, filters out repetitive, linguistically uninteresting material, and turns the web text into an offline corpus. The corpus can be improved iteratively by extracting new seeds from the first version of the corpus and repeating the procedure. Gatto particularly commends the fact that the researcher at will can manipulate the process at any point: how much data is wanted, from which domains (in Gatto's example, the search was limited to British and American academic sites, .ac.uk and .edu respectively), which of the suggested websites are actually included in the final corpus, and so on. This same service can be accessed online as WebBootCaT, through the website of the corpus query tool Sketch Engine. Gatto goes on with more, slightly repetitive, examples. However, her praise is convincing and not naïve. Finally, she explains how BootCaT and WebBootCaT can be used for compiling comparable corpora, "collections of texts similar in respect of genre,

topic, time span and communicative function" (p. 154), which can be useful to translators and lexicographers, and in Natural Language Processing; for example, for developing term banks and machine translation software and finding translational equivalents.

The last full-fledged chapter of the book, Chapter 6, combines corpus linguistics and Discourse Analysis in a brief case study on the meanings of *culture*. The chapter lists a variety of new, large-scale, online corpora from several languages, which are definitely worth taking a look at. The one employed in the case study is the 2-billion-word general purpose British English ukWaC corpus, which was collected, lemmatized, and annotated from 2005–2007 as part of the Web as Corpus Kool Ynitiative (WaCky). Gatto carefully explains the creation of the WaCky corpora, again highlighting the key challenges of (web) corpus compilation. Gatto argues, together with Baroni and Ueyama (2006), that many of these problems do not pertain to online corpora exclusively, but are only made more visible by the attempts to use the web as a corpus source because the web has allowed the collection of much larger bodies of text in a much shorter time than could be done using traditional corpus compilation methods.

Gatto's case study applies the web-based Sketch Engine tool for comparing the usage of the word *culture* in the classic British National Corpus and in the ukWaC, two of the many corpora that can be accessed with the Sketch Engine. It is an example of a new type of concordancers that allow the user to access various corpora online without having to download anything on their own computer, marking a change from corpus tools as "products" to corpus tools as "services" (Kilgarriff 2010). In addition to classic word lists and KWiC concordances, the Sketch Engine offers collocation lists with accompanying statistics, and so-called 'word sketches': "one-page automatic, corpus-based summaries of a word's grammatical and collocational behaviour" (Kilgarriff *et al*. 2004). Another interesting function is the comparison between sketches for two words, which measures preference for collocates and presents the data in coloured tables. Gatto looks at word sketches, concordances, and collocates of *culture* in the BNC and the ukWaC and suggests several possible changes in the meaning of the word. The investigation is, out of necessity, very cursory and because of this it does not necessarily do full justice to the Discourse Analytic method, but it does demonstrate some of the potential in combining the corpus linguistic approach with Discourse Analysis. The two main conclusions of the chapter are that these new web corpus tools change our way of "conceiving corpora", but the computer will nonetheless never analyse the data on our behalf (p. 202–203).

The last chapter touches very briefly on a recent development which is starting to gain more and more significance in research in and on the Internet:

the rise of the participatory 'Web 2.0', where content is, to an increasing degree, produced by users themselves. Gatto describes the main characteristics of the change, and then takes a brief look at two instantiations of Web 2.0 where she sees particular potential for corpus purposes: the collectively authored wiki-encyclopaediae as a multilingual corpus, and "cloud computing", where resources are distributed among numerous servers, possibly fostering "collective intelligence" among researchers (p. 209). Unfortunately, this is all Gatto has to say on the topic of Web 2.0 at this point. Since web language is explicitly left out of the scope of the book, Gatto does not engage the challenges of researching social media texts, such as ethical questions or non-standard orthography.

In the short conclusion, Gatto returns to the theoretical implications of combining the corpus approach with the web "as a huge self-generating collection of authentic text in machine-readable format" (p. 211). As final words, she stresses how the web as/for corpus studies should be seen as an alternative or supplement to, rather than as competition for 'traditional' corpus linguistics.

All in all, Maristella Gatto's *Web as corpus* is a readily understandable, pleasantly written review of the main developments of the interface of corpus linguistics and the World Wide Web. It combines practical and theoretical interests skillfully, and looks both to the past and to the future, not utopistically but with cautious optimism. The book is full of, but not heavy with, well discussed examples and illustrative figures. The tools and methods presented are potentially useful to language professionals in various fields – in particular to translators, lexicographers, and language teachers and learners, but assuredly to other linguists as well. The needs of novices are particularly taken into consideration, but more experienced researchers should also take a look at this work at the very least to make sure they are familiar with all the fascinating tools, methods, and corpora Gatto manages to meritoriously present to us in a mere 200 pages.

References

Baroni, Marco and Silvia Bernardini. 2006. *Wacky! Working papers on the web as corpus*. Bologna: Gedit. URL: http://wackybook.sslmit.unibo.it/.

Baroni, Marco and Motoko Ueyama. 2006. Building general- and special-purpose corpora by Web crawling. In *Proceedings of the NIJL International Symposium, Language Corpora: Their Compilation and Application*, 31–40. URL: http://home.sslmit.unibo.it/~baroni/publications/bu_wac_kokken_formatted.pdf.

Kilgarriff, Adam. 2010. Corpora by Web Services. *LREC workshop on web services and processing pipelines,* Malta. URL: http://trac.sketchengine.co.uk/attachment/wiki/AK/Papers/2010-K-WSPPmalta-CorporaByWeb-Sevices.pdf.

Kilgarriff, Adam, Pavel Rychly, Pavel Smerz and David Tugwell. 2004. The Sketch Engine. In *Proceedings Euralex*, Lorient, France, 105–116. URL: http://trac.sketchengine.co.uk/attachment/wiki/SkE/DocsIndex/sketch-engine-elx04.pdf?format=raw.

Leech, Geoffrey. 2007. New resources or just better old ones? The Holy Grail of representativeness. In M. Hundt, N. Nesselhauf and C. Biewer (eds.). *Corpus linguistics and the web*, 133–150. Amsterdam: Rodopi.

Wynne, Martin. 2002. *The language resource archive of the 21st century.* Oxford Text Archive. URL: http://www.lrec-conf.org/proceedings/lrec2002/sumarios/271.htm.