Roland Schäfer & Felix Bildhauer, *Web Corpus Construction* (Synthesis Lectures on Human Language Technologies 22). Morgan & Claypool, 2013. Pp. xv + 129.

Mats Wirén

Nordic Journal of Linguistics / Volume 37 / Issue 03 / December 2014, pp 457 - 463
DOI: 10.1017/S0332586514000304, Published online: 19 November 2014

REFERENCES

Berwick, Robert C. & Noam Chomsky. 2011. The biolinguistic program: The current state of its evolution and development. In Anna Maria Di Sciullo & Cedric Boeckx (eds.), *The Biolinguistic Enterprise: New Perspectives on the Evolution and Nature of the Human Language Faculty*, 19–41. New York: Oxford University Press.

Fitch, W. Tecumseh. 2010. *The Evolution of Language*. Cambridge: Cambridge University Press.

Hauser, Marc D., Noam Chomsky & W. Tecumseh Fitch. 2002. The faculty of language: What is it, who has it, and how did it evolve? *Science* 298, 1569–1579.

Lindblom, Björn. 1998. Systematic constraints and adaptive changes in the formation of sound structures. In James R. Hurford, Michael Studdert-Kennedy & Chris Knight (eds.), *Approaches to the Evolution of Language*, 242–264. Cambridge: Cambridge University Press.

McMahon, April M. S. & Robert McMahon. 2013. *Evolutionary Linguistics*. Cambridge: Cambridge University Press.

Tallerman, Maggie & Kathleen R. Gibson (eds.). 2012. *The Oxford Handbook of Language Evolution*. Oxford: Oxford University Press.

Roland Schäfer & Felix Bildhauer, *Web Corpus Construction* (Synthesis Lectures on Human Language Technologies 22). Morgan & Claypool, 2013. Pp. xv + 129. doi:10.1017/S0332586514000304

Reviewed by Mats Wirén

Department of Linguistics, Stockholm University, SE 106 91 Stockholm, Sweden.
mats.wiren@ling.su.se

With the advent of the World Wide Web around 1993, it gradually became apparent that an unprecedented source of linguistic data had come into existence for many languages. By 1998, 26 million unique URLs (uniform resource locators, roughly, web pages) were indexed by Google, itself launched in the same year. In 2000, the figure had risen to one billion (1,000,000,000) pages (Alpert & Hajaj 2008). A couple of years ago, estimates suggested that Google indexed about 40 billion pages (Fletcher 2013). In 2008, Google reported that they had identified the astonishing figure of one trillion (1,000,000,000,000) web pages, and that the number was 'growing by several billion pages per day' (Alpert & Hajaj 2008). By 1999, papers based on data obtained from the web were beginning to get published at major computational linguistics conferences (Kilgarriff & Grefenstette 2003). Broad acceptance of the usefulness of such data was manifested by the publication of a special issue on the web as corpus of the journal *Computational Linguistics* in 2003, the launch of the annual Web as Corpus (WaC) Workshop in 2005, and the founding of the ACL SIGWAC, the Special Interest Group of the Association for Computational Linguistics on Web as Corpus, in 2006.

Twenty years after the advent of the web, *Web Corpus Construction* arrives as the first introductory textbook on the steps required to compile a corpus from the web. Previously, an anthology (Hundt, Nesselhauf & Biewer 2007) and several overview articles had appeared (e.g. Kilgarriff & Grefenstette 2003, Fletcher 2013), including one to which the authors of the present book had contributed (Biemann et al. 2013). As the authors point out, the literature on this subject is quite diverse, involving techniques from search engines and data mining as well as computational linguistics; perhaps this is one reason why it took so long for a textbook like this to appear. Apart from the comprehensive introduction, a valuable contribution of the book is its collection of references to the vast literature on the subject; the bibliography comprises almost 15% of the book. The book also contains references to freely available, open-source tools for the different steps of web corpus construction, though it does not provide introductions to any of them.

The most obvious way of employing the web for linguistic research is to use a commercial search engine, like Google, Yahoo!, or Bing, or a system that post-processes results from these, such as WebCorp (Renouf, Kehoe & Banerjee 2007). An appealing feature of this approach is that it means directly using the (indexed) web as a corpus, but the arbitrariness of search engine counts and the opaqueness of ranking criteria make the interpretation of such results fundamentally problematic, if not impossible (Kilgarriff 2007). The scientific solution is therefore to compile a corpus of one's own by extracting documents from the web. There are different ways of doing this. For specific content, it may be sufficient to download a particular website according to predefined design criteria, much in a way that corresponds to the sampling of material for a traditional corpus (of, say, newspaper articles). Alternatively, if one wants to take advantage of the seemingly unlimited resources of the web, one can download a large-scale data set possibly consisting of billions of words. Although it is not stated quite clearly, the focus of *Web Corpus Construction* is on this latter approach.

There are three major ways in which the compilation of such a web corpus differs from that of a traditional corpus: (i) Sampling is realised through crawling, which uses the connectedness of the web to collect pages by recursively following links to new pages, typically starting with a set of well-connected initial pages (called the seeds). (ii) Much material needs to be filtered out after crawling. First, whole documents need to be discarded because of duplicates or lack of textual content. In one of the web corpora referred to by the authors (the DECOW2012 corpus in Figure 2.6 on page 20), 94% of the crawled documents were filtered out for these reasons. (DECOW2012 is the German corpus compiled in 2012 of the COW corpora collection, Corpora from the Web.) Secondly, lots of material within documents needs to be filtered out because of markup, navigational menus, etc. The final yield rate, that is, proportion of the size of data retained in the final corpus and the size of downloaded data, may be just a few percent or less (Suchomel & Pomikálek 2012).

(iii) Whereas traditionally sampled corpora are carefully balanced in terms of content and genre, the exact composition of a web corpus has to be inferred a posteriori, since typically all material that is encountered during crawling is initially included.

The book defines (p. 7) a 'web corpus' as a static set of documents (a snapshot) collected from the web. The limitation to a static set is motivated by the goal of reproducibility; a side-effect of this is that the added complexity of incremental crawling does not have to be addressed.

Chapter 1, 'Web corpora', briefly recapitulates the motivations for web corpora in comparison with traditionally sampled corpora: virtually no cost, almost arbitrary size, and great variability of the language exhibited, including certain genres (such as blogs) that are only available on the web. To give the reader an idea of size, the WaCky corpora for English, French, German and Italian (Baroni et al. 2009) comprise around 1–2 billion words each, and the latest COW corpora being compiled for English, French, German, Spanish and Swedish are in the order of 10 billion words each (Schäfer & Bildhauer 2012; see also http://hpsg.fu-berlin.de/cow/). Although these corpora are all available on request, copyright and redistribution of web corpora are not legally uncontroversial. Another disadvantage of web corpora is that their noisiness – deviations such as non-standard spelling and faulty punctuation from the kind of standard language usually dealt with in corpus linguistics – may render them harder to use than traditional corpora (a problem which is discussed in Chapter 3). To these circumstances mentioned in the book might be added the scarcity of 'metadata' for web documents (beyond the Internet domain), such as facts about the author, intended target audience, purpose, etc.

The remaining chapters are organised according to the natural progression of corpus construction. Chapter 2, 'Data collection', begins with a discussion of the structure of the web, including segments that are problematic from the point of view of corpus construction, such as dynamic web pages. This is followed by a detailed description of the basic steps of crawling: seed URLs (the pages to start with), constraints on the crawler with respect to content (domains, file types, languages, etc.), politeness settings (such as the frequency with which the crawler sends requests to a host), and search strategies. Finally, approaches for avoiding bias in sampling, and for guiding the selection of material (focused crawling), are discussed. In the beginning of the chapter, the authors note that 'data collection for web corpus construction has received the least amount of attention from the linguistic community' (p. 7), and that this chapter is therefore the least practical of all. In particular, only a single paragraph (on page 22) deals with general software tools for crawling, where the free products Heritrix and Nutch are mentioned. Here one wonders if the argument could not have been turned around: it seems that this lack of attention of linguists would have provided a good motivation for including some practical introduction related to crawlers. It is claimed that Heritrix is relatively easy to configure, and this system appears to be very well suited to the kind of large-scale data collection that is the

focus of the book. As touched on above, however, many linguists may be primarily interested in capturing particular websites related to their specific research problems, and then it is not clear if Heritrix or Nutch are the most convenient options.

Chapter 3, 'Post-processing', describes non-linguistic cleaning methods for web corpora. This involves stripping of markup (mainly HTML), character set conversion (since different web pages have different character encodings), language identification (for the purpose of filtering out documents not in the target language), detection of non-text documents, detection of duplicate and near-duplicate documents, and boilerplate removal. The term 'boilerplate' refers to templates such as menus, headers, footers, navigational elements and advertisements that are repeated over several pages and that are usually discarded for the purpose of corpus construction. Several tools that perform one or several of these tasks are mentioned in the different sections.

Chapter 4, 'Linguistic processing', describes the basic steps of linguistic annotation, namely, tokenisation, part-of-speech tagging and lemmatisation, particularly from the point of view of noisy data. After a brief overview of the general methods, the concept of noise in web corpora is discussed. This is taken to include misspellings, tokenisation errors, remaining non-words and foreign-language material. It is argued that a major problem with noise is that figures concerning lexicon size may become misleading; in the DECOW2010 web corpus, one half of the hapax legomena (words occurring only once) results from such noise. The book then pinpoints two cases in which noise causes problems for tokenisation of web texts: missing whitespace (horizontal space characters), particularly at sentence ends, and emoticons (emotion icons), a salient feature in several web genres. This is followed by a discussion of the ways in which noise provides obstacles to part-of-speech tagging and lemmatisation, concluding with several useful recipes for circumventing these problems. Next, the problem of spelling variation and orthographic normalisation is discussed, and a final section briefly describes software tools for linguistic postprocessing. For readers interested in Scandinavian languages, it should be noted that there is a freely available, state-of-the-art part-of-speech tagger trained so far for Swedish and Icelandic, which has already been applied to Swedish web text, namely, Stagger (see Loftsson & Östling 2013, also Östling 2013). On a related note, linguistic processing beyond part-of-speech tagging is not covered in the book, but one technique which has gained enormously in terms of robustness during the last decade is syntactic analysis. As an example of current practice, MaltParser (Nivre et al. 2007) has been used for syntactic annotation of the entire Korp corpus collection of Språkbanken (http://spraakbanken.gu.se/korp/), the major part of which is web corpora.

Finally, Chapter 5, 'Corpus evaluation and comparison', discusses ways of assessing the technical quality and composition of a web corpus. First, two surface indicators of potential quality problems are described, namely, abnormal distribution

of word and sentence lengths, and abnormal duplication of large *n*-grams and sentences. Next follows a very useful discussion on how to compare a web corpus with other corpora, either compiled from the web or created using traditional sampling methods. In this discussion, the authors draw on Kilgarriff (2001) and argue that rather than using hypothesis testing for the purpose of trying to determine if two corpora correspond to samples from the same population, it is more informative to use the test statistic (for example, χ^2) to assess the relative similarity between corpora, much like the use of collocation measures. A related way of characterising a corpus is to use lists of keywords, for example, ratios of relative frequencies of words. Next, approaches to evaluation of web corpora are discussed. Here, the focus is on extrinsic evaluation, that is, an assessment of the usefulness of a web corpus based on a particular application, such as finding collocations or using it for training of tools such as part-of-speech taggers. Finally, the problem of determining the composition of the web corpus in terms of genres, text types, etc. is discussed. It is pointed out that there is no such thing as an established inventory of web genres; nevertheless, a couple of approaches for determining composition are described, and the notions of balance and representativity of web corpora are discussed.

Who is the intended reader of this book? My impression from the Preface is that the model reader is anyone with or without programming skills who would like to compile a web corpus using available tools or their own tools. The book thus seems most suitable to (computational) linguists interested in understanding the details about the processing chain for a web corpus, either as a prerequisite for adapting and configuring existing tools, or as a reference for the development of new tools. In the former case, however (and as mentioned above), some practical introduction to crawling tools would have been a worthwhile addition, perhaps also aimed at the situation when the goal is not a large-scale corpus. As for linguists interested in analysing web corpora but not directly involved in their construction, some of the technical detail may not be accessible or relevant. In this case, however, parts of the book will still provide a useful background for the purpose of understanding the limits and possibilities of the methodology.

Is there anything apart from the inclusion of a practical introduction to crawlers and of syntactic parsing that might have increased the value of the book? Yet another topic which I would have liked to see discussed in this volume is legal aspects of web corpus construction, especially since redistribution is clearly a key issue in many of the projects referred to in the book. A problem in so doing is that regulations and legal circumstances vary considerably across countries, but it would at least have been useful to have a summary of common practice employed to circumvent current problems. One example of such practice is re-shuffling the sentences of the corpus to make it impossible to reconstruct; another is allowing copyright owners to 'opt out', by not notifiying them about their material being included in a corpus, but allowing them to request to have it excluded if they find out about it, as in the WaCky project

(Baroni et al. 2009). Finally, one other feature which would have increased the value of the book is a proper index (currently, only page numbers for the entries in the bibliography are indicated).

Summing up, I found that this book successfully fills a pedagogic gap in the corpus-linguistics literature. It is well-structured, coherent and well-written, and I enjoyed reading it. At 129 pages, the book is relatively short, yet all the relevant steps of web corpus construction are covered in a reasonably comprehensive manner. Moreover, as noted above, the overview of the literature is a valuable contribution in itself. As a bonus, parts of the book (especially Chapters 4 and 5) will be relevant to corpus construction in general, given that several of the methods overlap and that noisy data are not unique to the web. The book has a companion website, http://sites.morganclaypool.com/wcc, which includes slides from the authors' ESSLI (European Summer School in Logic, Language and Information) 2012 course 'Building large corpora from the web' and links to web corpora and software (tools for crawling, cleaning, corpus indexing and linguistic processing), many of which are supplementary to the book.

REFERENCES

Alpert, Jesse & Nissan Hajaj. 2008. We knew the web was big . . . http://googleblog.blogspot.se/2008/07/we-knew-web-was-big.html.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi & Eros Zanchetta. 2009. The WaCkyWide Web: A collection of very large linguistically processed webcrawled corpora. *Language Resources & Evaluation* 43, 209–226.

Biemann, Chris, Felix Bildhauer, Stefan Evert, Dirk Goldhahn, Uwe Quasthoff, Roland Schäfer, Johannes Simon, Leonard Swiezinski & Torsten Zesch. 2013. Scalable construction of high-quality web corpora. *Journal for Language Technology and Computational Linguistics*, 28(2), 23–59.

Fletcher, William H. 2013. Corpus analysis of the World Wide Web. In Carol A. Chapelle (ed.), *The Encyclopedia of Applied Linguistics*, vol. 3, 1339–1347. Oxford: Wiley-Blackwell.

Hundt, Marianne, Nadja Nesselhauf & Carolin Biewer (eds.). 2007. *Corpus Linguistics and the Web*. Amsterdam: Rodopi.

Kilgarriff, Adam. 2001. Comparing corpora. *International Journal of Corpus Linguistics* 6(1), 97–133.

Kilgarriff, Adam. 2007. Googleology is bad science. *Computational Linguistics* 33(1), 147–151.

Kilgarriff, Adam & Gregory Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Computational Linguistics* 29(3), 333–347.

Loftsson, Hrafn & Robert Östling. 2013. Tagging a morphologically complex language using an Averaged Perceptron Tagger: The case of Icelandic. *19th Nordic Conference of Computational Linguistics* (NODALIDA), 105–119. Linköping: Linköping University Electronic Press.

Nivre, Joakim, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov & Erwin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering* 13, 95–135.

Östling, Robert. 2013. Stagger: An open-source part of speech tagger for Swedish. *Northern European Journal of Language Technology* (NEJLT) 3, 1–18.
Renouf, Antoinette, Andrew Kehoe & Jayeeta Banerjee. 2007. WebCorp: An integrated system for web text search. In Hundt et al. (eds.), 2007, 47–67.
Schäfer, Roland & Felix Bildhauer. 2012. Building large corpora from the web using a New Efficient Tool Chain. *The Eighth International Conference on Language Resources and Evaluation* (LREC), Istanbul, Turkey, 486–493.
Suchomel, Vít & Jan Pomikálek. 2012. Efficient web crawling for large text corpora. *The Seventh Web as Corpus Workshop* (WAC), Lyon, France, 39–43.

Jon Sprouse & Norbert Hornstein (eds.), *Experimental Syntax and Island Effects*. Cambridge: Cambridge University Press, 2013. Pp. x + 421.
doi:10.1017/S0332586514000316

Reviewed by Fredrik Heinat

Department of Languages, Linnæus University, 351 95 Växjö, Sweden.
fredrik.heinat@lnu.se

Even though dependencies in language are often very local, some types of dependencies can span infinitely many clauses. One such type of dependency is filler–gap dependencies. A filler–gap dependency is the dependency we find between a *wh*-word, or any other fronted constituent (the filler) and its thematic position (the gap), as in: *Which book did Mary say that Liz claimed that Emma read* ___. At least since Ross (1967), it has been known that there are certain structures into which filler–gap dependencies cannot so easily be formed. These structures are called islands. The present volume consists of 16 chapters on various aspects of island phenomena. The book is divided into two parts. The first part, 'Global issues in the investigation of island effects' (Chapters 2–6), is concerned with the two main approaches to islands. One approach is to see island effects as a consequence of (universal) syntactic constraints. The other approach is to view the effects as a consequence of an overload of general processing processes (such as working memory). The second part, 'Specific issues in the investigation of island effects' (Chapters 7–16), consists of chapters dealing with mostly previously published studies that have made use of controlled acceptability judgements. There is not enough space in this review to treat every chapter in detail, but the chapters from Part I are treated somewhat more extensively since they deal with more general matters than the chapters in Part II.

In Chapter 1, 'Experimental syntax and island effects: Toward a comprehensive theory of island effects', Jon Sprouse & Norbert Hornstein state that the aim of the present volume is to examine what experimental syntax can tell us about island effects. The term effect refers to the difference in acceptability rating of sentences containing dependencies into islands and sentences with dependencies into