which need to be understood, covering important areas of colour management and image quality, which are often overlooked by other text books in the field.

The final section, 'Applications', considers the digitization of particular types of object, such as architectural sites, stained glass windows, easel paintings, and sculpture, and looks at how advanced technologies can aid in the capture and delivery of particular types of digital images. The book finishes with an overview chapter on 'Research Policy and Directions', providing a stimulating overview of future possibilities for the application of imaging in this arena.

The range of topics covered within the text presents applications of imaging technologies beyond the usual scope of the library, archive, and museum sectors, for example, the use of digital imaging in commercial art galleries or sites of architectural interest, is often overlooked. As such, *Digital Heritage* shows how central digital imaging has become to both public and private sectors, and how increasingly reliant we have become on imaging technologies. By encouraging information professionals to develop an understanding of the underlying principles behind digital imaging, and grasping the pertinent issues to the application of these technologies to the cultural and heritage sectors, this text encourages us to see beyond traditional institutional boundaries, and consider how digital imaging of documents, images, and artefacts relates to the use of images in wider culture.

This informative, varied, and well-illustrated text should be viewed as an essential companion to the managerial texts on the instigation of digitization projects, demonstrating the strengths of digitization for the cultural and heritage sectors, whilst illustrating and explaining at first hand how the application of imaging technologies to complex, historical artefacts can increase our understanding of, and access to, the objects themselves.

References

Besser, H. (2003). *Introduction to Imaging. Revised edition*. Santa Monica, Getty Information Institute: Getty Trust Publications <http://www.getty.edu/research/conducting_research/standards/introimages/> (last accessed 11 February 2008).

Deegan, M. and Tanner, S. (2002). *Digital Futures: Strategies for the Information Age*. London: Library Association Publishing.

Hughes, L. (2004). *Digitizing Collections: Strategic Issues for the Information Manager*. London: Facet Publishing.

Kenney, A. R. and Reiger, O. Y. (2000). *Moving Theory into Practice: Digital Imaging for Libraries and Archives*. Mountain View CA: Research Libraries Group.

Lee, S. (2002). *Digital Imaging, A Practical Handbook*. London: Facet Publishing.

Melissa Terras, School of Library, Archive and Information Studies, University College London, UK

doi:10.1093/llc/fqn002
Advance Access Published on 15 February 2008

Corpus Linguistics and the Web.

Marianne Hundt, Nadja Nesselhauf and Carolin Biewer (eds). *Amsterdam/New York: Rodopi, 2007. 305 pp. ISBN: 90-420-2128-4. $94/€65 (hardback)*.

New perspectives in corpus linguistics explore the use of the web as a corpus or for corpus building. The enormous amount of data on the web can be a valuable resource for research concerning lexical innovations, morphological productivity, or other topics for which even the available mega-size corpora, like the 100 million words British National Corpus (BNC), are not large enough. Corpus study of outer-circle varieties of English becomes feasible since the web contains authentic, non-edited texts in e.g. Samoan English. The web has also given rise to new text types—like e-mail and weblogs—which are not represented in traditional corpora. Moreover, the web is subject to continuous change, while traditional corpora are static and quickly out of date. Web language—so-called *weblish* or *netspeak*—may have an influence on language change. These phenomena can be investigated by using the web for corpus linguistics.

The volume *Corpus Linguistics and the Web* contains a selection of papers presented at the 2004 symposium on *Corpus Linguistics—Perspectives for the Future*, held at the *Internationales*

Wissenschaftsforum (IWH) in Heidelberg, Germany. A number of invited papers was commissioned later. The papers are organised in four sections: the first section focuses on advantages and problems with *Web as Corpus* (WaC) and *Web for Corpus* (WfC), section two presents corpora compiled from the internet, section three adds some critical voices, and section four introduces some applications for language variation and change.

In the first chapter, Lüdeling, Evert and Baroni give an overview of the state of the art in using web data for linguistic purposes, and discuss advantages and limitations of the different approaches. Using a commercial search engine (SE) allows for fast data collection, but pre- and/or post-processing is needed to refine query results (e.g. *WebCorp*, *KWiCFinder*). Building an SE for linguists may be the solution for abstract or linguistic querying, but this would require 'major computational resources and very serious, coordinated, high-efficiency programming' (p. 20) and raise legal issues. WfC exhibits a number of advantages over WaC: it allows for reproducibility, higher search accuracy, and incorporation of metadata.

Fletcher (Chapter 2) introduces *KWiCFinder*, a pre- and post-processing tool for the Altavista SE for Wa/fC. This software tool enables more abstract querying—like the so-called 'tamecards' *s[iau]ng[,s,ing]* (p. 34), which query for all forms of the verb *sing*—and generates concordance reports with keyword-in-context (*KWiC*) information. Fletcher also discusses the 'pitfalls of ''webidence'' ' (p. 25): the figures of SEs are not to be trusted and verifiability—the idea that other researchers should be able to duplicate your research and test your results—is not possible with WaC.

A new architecture for Web*Corp*, a different tool that allows retrieval of raw and linguistic data from the web, is proposed by Renouf, Kehoe, and Banerjee in the third chapter. The new W*ebCorp Linguistic Search Engine* will build up the corpus incrementally, integrate 'neologisms, summaries, document similarity measures, domain identification and so on' (p. 61) and improve speed, statistics, and search.

The second part of this volume concerns the compilation of corpora from the web. Hoffmann (Chapter 4) presents a corpus of CNN transcripts automatically retrieved from the CNN website. Named entities are normalised, so that *Soledad O'Brien*, *O'Brien*, and *CNN anchor* refer to the same person. By storing the data in databases, Hoffmann is able to perform a case study on the use of *so not*. Claridge (Chapter 5) collects data from message boards from Great Britain, USA, Australia, and Canada, in order to study regional varieties of English in forum language. A 3.7 million word corpus of data from Google categories *Home* and *Science* is selected by Biber and Kurjian (Chapter 6). Their goal is to build a taxonomy of web registers and text types by performing multi-dimensional analysis. One approach relies on types defined on the web, while a second approach takes into account web text types that are linguistically defined. Biber and Kurjian show that SE categories are not well defined for linguistic research.

In section three, we find some critical voices concerning the use of web data for corpus linguistics. Leech (Chapter 7) states that corpus linguistics should not 'neglect to improve and refine the resources and methods we already have' (p. 133). The quest for the 'Holy Grail of representativeness' (p. 134) should be put into perspective: no corpus linguist will ever be able to collect all data in the textual universe. BNC and Brown Corpus are said to be *balanced* corpora, but this would imply knowledge of the proportion of every genre in the textual universe, according to Leech. He proposes two solutions: an estimation of text usage based on external and internal criteria, and the collection of *comparable* corpora, two or more corpora that differ in only one parameter. Corpus linguists can disregard the notions of representativeness, balance and the idea of a sample corpus when using WaC.

According to Kennedy (Chapter 8), the British National Corpus (BNC) is one of the underexploited resources Leech (Chapter 7) mentioned. He claims that the BNC overcomes some of the problems WaC has difficulties dealing with: representativeness, authorship, copyright issues and the large amount of data. In the description of English and exploration of language learning and teaching, BNC can be a very useful resource.

The fourth part of this volume focuses on case studies of Wa/fC. In Chapter 9, Rosenbach

compares Google and *WebCorp* frequencies for the use of '*s*-genitives (e.g. *driver's licence*) versus noun collocations (*driver licence*). Absolute and especially relative frequencies are dramatically different since the SE does not discriminate properly between forms with and without '*s*, which is one of the reasons why *WebCorp* was built. Rohdenburg (Chapter 10) studies the interaction between SE data and a large corpus collection of newspaper articles to find that WaC has several advantages over traditional corpus analyses. It allows for fast succession of studies, examination of fine-grained problems, and exploration of issues rare or absent in formal written corpora. In Chapter 11, Mondorf compares the use of synchronic and historical corpora, with internet data retrieved by means of a SE. The focus is on morpho-syntactic variation in the comparative alteration (*fresher* vs. *more fresh*). Mair (Chapter 12) studies change and variation in present-day English in closed corpora and the web. Despite uncertainty concerning the amount and quality of web data, WaC is said to be successful. Hundt and Biewer (Chapter 13) hypothesize that Australian English and New Zealand English influence varieties of English in the South Pacific and East-Asia. A manually selected corpus of newspaper articles from the web is however not able to confirm the central hypothesis. In Chapter 14, Anderwald compares data retrieved from the web—via a SE or *WebCorp*—to the Freiburg English Dialect Corpus (FRED) for a case study on non-standard past tense forms (e.g. *He rung the bell*). She concludes that the SE is linguistically less accurate and reported frequencies are not reliable, but nevertheless documents authentic, unedited language. The last paper in this volume, by Nesselhauf (Chapter 15), focusses on a diachronic study of *will* and *shall* with future reference in ARCHER (A Representative Corpus of Historical English Registers) and WebFict, a corpus of novels collected from Project Gutenberg. Nesselhauf sees great potential in WebFict for diachronic analysis.

This book offers a clear overview of a new perspective in corpus linguistics: the use of the web. The first section gives a realistic idea of the current state-of-the-art (*KWiCFinder*, *WebCorp*), the advantages and limitations of Wa/fC. In the second section, the potential of Wa/fC is clearly demonstrated in studies using large-scale corpora collected from the web. The critical voices in the third section were a welcome surprise—or even relief—after a few papers about problems and limitations. The central idea there could be reformulated as 'Why should we bother using the web for corpus linguistics when we have corpora like the BNC?'. After this climax, the papers of the fourth section seem in general to provide fewer new insights. Surprisingly, most of these papers do not take into account the limitations of Wa/fC formulated in the initial chapters. Some of the papers use simple Google search, acknowledge that frequencies reported by SEs should not be trusted, but still draw conclusions from them. Some other critical observations can be made concerning Chapters 11 and 14. Mondorf (Chapter 11) compares results from Google search with a historical corpus on fiction and a synchronic newspaper corpus in a study on the comparative alternation. Apart from the unclarity about the type of data retrieved from Google, Mondorf disregards the genre and register differences between novels and newspaper articles, which implies that they are incomparable. A similar comment can be made for the paper by Anderwald (Chapter 14), which compares weblogs and webforums with BBC and government data.

Overall, this book is a valuable resource for anyone interested in applying the enormous amount of data on the web to linguistic research. Especially the first three parts offer an objective view on advantages, problems, and limitations. Although the fourth part is technically less convincing, it demonstrates the array of applications of the web for corpus linguistics.

Kim Luyckx, University of Antwerp, Belgium

doi:10.1093/llc/fqn001
Advance Access Published on 19 February 2008