DE GRUYTER
OPEN

Review in ICAME Journal, Volume 38, 2014, DOI: 10.2478/icame-2014-0017

Anneli Meurman-Solin and **Jukka Tyrkkö** (eds.). *Principles and practices for the digital editing and annotation of diachronic data* (Studies in Variation, Contacts and Change in English 14). Helsinki: Varieng. http://www.helsinki.fi/varieng/series/volumes/14/. 2013. Reviewed by **Anne Gardner**, University of Zurich.

Introduction

Over the past decades the technologies available for representing texts from historical stages of a language in electronic format have advanced to such an extent that it is now feasible to supply digitally edited texts with annotation concerning not only the language, but also material aspects of the original. With this collection of articles Meurman-Solin and Tyrkkö illustrate some of the methodologies which have been developed in order to adapt modern technologies to historical data. The editors propose that, for a fuller understanding of the language of a text, its context and non-linguistic features such as layout, script type and abbreviations have to be taken into consideration as well, and that it is consequently important to adequately represent such features through the annotation of a corpus. This approach had been advocated by the editors, amongst others, at the *Helsinki Corpus Festival* in autumn 2011 as one of the future directions for English historical corpora.

The publication is divided into four parts, the first of which offers a description and critical assessment of the aims, development and uses of three important corpora or family of corpora. These are the Helsinki Corpus of English Texts (HC), published in 1991, and two of its off-shoots developed later on at the Research Unit for Variation, Contacts and Change in English (VARIENG, established in 1995), namely the family of corpora based on the Corpus of Early English Correspondence (CEEC) and the Corpus of Early English Medical Writing (CEEM). Parts II and III discuss the relevance of visual features in correspondence and trial proceedings, as well as the significance of paratextual characteristics of printed title-pages for a context-based linguistic analysis. From the point of view of four different projects, Part IV presents methodological and theoretical principles underlying the encoding of linguistic and non-linguistic features of manuscript and printed sources.

Part I: On the evolution of three major corpus projects in English historical linguistics

The articles of this section offer interesting information relating to the HC, the CEEC family of corpora and the CEEM, more details of which can be found on the CoRD website (http://www.helsinki.fi/varieng/CoRD/index.html), as well as a brief review of the use of the corpora by the international research community. Furthermore, the authors evaluate each corpus in terms of relevance, representativeness, quality and authenticity of the texts, as well as depth of the coding of language-external variables. **Rissanen** and **Tyrkkö** present some of the editorial approaches to, for instance, the text selection and subperiodisation of the original HC, as well as motivations behind the conversion project resulting in an XML version which preserves the original corpus in a more modern format, while also incorporating corrections in the parameter encoding and the corpus texts themselves. **Nevala** and **Nurmi** show that the CEEC corpora have been valuable in particular for stratificational and interactional sociolinguistics, as well as socio-pragmatics, and note that efforts are being made to implement spelling normalisation with the help of the VARD2 tool (http://www.comp.lancs.ac.uk/~barona/vard2/). **Taavitsainen** and **Pahta** report that work on CEEM has revealed that the beginnings of scientific writing in English lie in the late fourteenth rather than in the early eighteenth century. They illustrate that it is possible to trace the (non-linear) development from medieval scholasticism to early-modern empiricism and observe a move away from the reliance on received knowledge. CEEM is therefore a valuable resource not only for the linguistic analysis of English in a particular register, but also for medical and cultural studies concerned with scientific writing.

Part II: Material features in manuscripts. Correspondence and trial proceedings in focus

The four contributions in Part II focus on paratextual features offering information on the structure of letters and depositions. In her article on visual prosody, **Meurman-Solin** showcases to what extent normalisation and standardisation practices of editions can distort manuscript realities. Her discussion is based on letters re-edited for the Corpus of Scottish Correspondence 1500–1715 (CSC), which aims at providing a diplomatic transcription of the manuscripts. She demonstrates the importance of annotating features of visual prosody such as punctuation, spacing and marked character shapes, which offer information relevant for the analysis of syntax, discourse and text structure. **Walker** and **Kytö** illustrate features of layout, such as empty space, alignment or indentation, and handwriting (e.g. font changes, letter shapes) worth studying in An Electronic

Text Edition of Depositions 1560–1760 (ETED), not all of which are encoded (yet) in the corpus. They show that layout practices differ between church depositions on the one hand, and criminal court depositions on the other: they were more varied in criminal court depositions, which were mostly recorded in diverse localities and then sent to a court, than in church depositions, which were compiled in one place and bound together, involving fewer scribes. Although **Sairio** and **Nevala** do not touch on actual annotation principles and practices for digital editions, their contribution is nevertheless a useful addition to this volume in that it stresses the importance of layout features pertaining to the structure of letters – the use of space in salutations and subscriptions in particular – for the assessment of style and politeness. Letter-writing manuals indicate that certain rules of spacing should be followed to honour the social distance between sender and addressee, but in four case studies of letters written by members of the Bluestocking circle it is shown that such norms could be disregarded if the relationship between sender and addressee was very close. Sairio and Nevala's article is particularly important as it provides the reader with the necessary background on why certain layout features are used in letter-writing in order to contextualise Meurman-Solin's almost purely descriptive account of such structural features in letters from the CSC. A large body of images serves to exemplify, for instance, that in letters from the early and mid-sixteenth century the closing formula and signature were typically separated from the text body by a wide space to signal deference to the addressee, or that towards the later period initial and final formulae tended to be allowed more space than previously.

Part III: Paratextual properties in early printed title-pages

Invoking the concept of discourse community, **McConchie** reflects on the roles of author, printer, publisher and consumer in the production and reception of title-pages. Using examples from the mid-sixteenth to the early seventeenth century, he thoroughly explores in how far elements of title-pages, including capitalisation, small capitals, italics, embellished characters, point size, colour, or graphic representations, may have linguistic function and serve to support its illocutionary force. Further linguistic layers are created by later additions made to title-pages, for instance those indicating ownership (signature, date and place of acquisition etc.). **Ratia** carefully examines the title-pages of fifteen plague treatises from the Stuart period in order to identify textual labels which function as genre markers and determine in how far labels such as 'treatise', 'directions', 'discourse', 'queries' or 'prayer' correlate with the contents of the texts. Plague treatises from this time are found to fall into three groups, i.e. religious, religio-

medical and non-religious (i.e. medical), eight of the selected works belonging to the second category. Even in non-religious texts, quotations from the Bible may appear on the title-page; however, religious content is generally not strongly promoted on title-pages, neither through textual labels nor through visual highlighting.

Part IV: New approaches to digital editing

Building on a discussion of the various types of abbreviations for Latin and English words and their history, as well as their taxonomisation in handbooks and their treatment in printed editions, **Honkapohja** presents in useful detail the system he adopted for encoding abbreviations in his digital edition of *The Trinity Seven Planets*. A particular advantage of XML-based tagging is that abbreviations can be represented both in the original and in an expanded version, while it remains transparent which changes or expansions are made by the editor. Using different attributes, it is possible to mark the degree of certainty behind such expansions as well as different types of abbreviations. **Meurman-Solin** critically addresses the issue of how features of visual prosody, or 'indirectly linguistic' features, could be taxonomised in a way which acknowledges variation across time and space, for instance, or between different discourse communities, ranging from ideolectal to grouplectal and community-wide practices. Arguing against a simple polarisation between 'default' and 'marked' as well as a frequency-based taxonomy, she proposes that features could be tagged with strings of co-ordinates through which 'categorial fuzziness and polyfunctionality' can be taken into account – a co-ordinate is defined as 'a long-diachrony-based linguistic property'; an illustration of this intriguing approach would have been helpful. On the basis of examples from the Lampeter Corpus, **Claridge** reviews in how far visual features may be beneficial to linguistic analysis and to what degree they are encoded in this corpus. Concerning title-pages she convincingly argues that layout, as well as type and font variation aim at emphasizing particular words or phrases; if only selected aspects are encoded, such as occurrence of italics alone, the original emphasis can easily be misrepresented in the corpus. Another interesting finding is that the use of black-letter merits annotation since it often serves a particular function and may reflect the intention of the author rather than the editor. Claridge provides excerpts on the composition of title-pages and typography from a late-seventeenth-century source which also serve as illuminating background information for the articles in Part III in particular; adding cross-references to Claridge might therefore have been advantageous. Investigating 95 title-pages of works associated with the apothecary Nicholas Culpeper, **Tyrkkö**, **Marttila** and **Suhr** introduce the annotation system applied

in the preparation of their corpus and successfully demonstrate how attention to paratextual features can assist in identifying distinct printing-house styles and their strategies for marketing publications written by or associated with Culpeper. For instance, in order to emphasize Culpeper's name, printers belonging to Strand B use larger type sizes, also in relation to the main title, while those from Strand A prefer to set the name in a different type; use of black-letter is argued to highlight Culpeper's importance as a provider of medical texts in the vernacular at a time when Latin was still predominantly used in this domain.

Evaluation

By opting for an electronic publication, Meurman-Solin and Tyrkkö have chosen an ideal medium for their purpose. Unrestricted by constraints of space and related issues which are usually encountered in the case of a printed book, the present volume, in particular Parts II and III, can boast a wealth of large reproductions of original letters and early prints which illustrate the points made by the authors. Similarly, Honkapohja's contribution offers a substantial number of close-ups of manuscript abbreviations which are accompanied by their respective representations in XML. Here, as in the articles by Rissanen and Tyrkkö, as well as Tyrkkö, Marttila and Suhr, XML encoding is provided in colour, which increases reader-friendliness significantly. Other technical niceties proffered by the electronic medium include the possibility of 'scrolling' through letter images to view all pages.

This timely collection of articles, some of which might have benefitted from another round of proof-reading, will be of interest to researchers working in the field of corpus compilation and digital editing who need to adapt modern technologies and principles to historical data. Introduction and articles alike offer a transparent and critical account of relevant approaches and methodologies, besides highlighting current research trends in general. In addition, scholars are encouraged to analyse (digitally) edited texts with a mind to the context in which they were produced and to physical aspects of the originals as non-linguistic features are shown to provide important additional information relevant for the linguistic analysis of the texts. Overall, the volume promotes a *rapprochement* of linguistics and material text studies within the field of digital humanities. The authors of the articles, many of whom are associated with the VARIENG research unit, can all draw on personal experience with regards to corpus compilation and digital editing, and are consequently able to provide valuable insights into these areas. The volume may therefore serve as inspiration and provide guidance especially to those who are planning or are in the early stages of their own corpus compilation or digitisation project.