Review

Early Modern English Medical Texts. 2010. *Corpus* [CD-ROM]. Compiled by Irma Taavitsainen, Päivi Pahta, Turo Hiltunen, Martti Mäkinen, Ville Marttila, Maura Ratia, Carla Suhr, and Jukka Tyrkkö. Software by Raymond Hickey. *Corpus Description and Studies*. Edited by Irma Taavitsainen and Päivi Pahta. Amsterdam/Philadelphia: John Benjamins.

Since the advent of the *Helsinki Corpus of English Texts* in the 1990s, compilers of English historical corpora have moved increasingly toward providing "short and fat" electronic text databases instead of "long and thin" (p. 7). These "short and fat" corpora commonly focus on a particular domain or genre and are more limited in temporal scope; this setup enables more in-depth studies of domain-specific language as well as investigations of patterns underpinning the more general trends found in the "long and thin" corpora. *Early Modern English Medical Texts* (henceforth EMEMT) is a very welcome contribution to this trend. It provides a continuation of *Middle English Medical Texts*, which was released in 2005 (Taavitsainen, Pahta, and Mäkinen), and is projected to be followed by a Late Modern English collection (p. 2). Together with the book that introduces the corpus and illustrates some of its possible uses, EMEMT constitutes a rich source for explorations of the connection between language and society, and presents new developments in the compilation and presentation of historical corpora.

The EMEMT package consists of several components: the corpus itself, the book with descriptions and studies (which includes sections by scholars other than the compilers), and the software that accompanies the corpus. I will begin by discussing general features of the corpus. Although I will draw on and refer to chapters in the book in this discussion, I will return to a description of the book as such. Finally, I will explore the technical aspects of the EMEMT, including software, coding, and presentation.

EMEMT consists of two million words of extracts from medical texts from the period 1500 to 1700. The texts are divided up into six main categories: 1) General treatises and textbooks; 2) Treatises on specific topics (further subdivided into Texts on specific diseases, Texts on plague, etc.); 3) Recipe collections and materia medica; 4) Regimens

and health guides; 5) Surgical and anatomical treatises; and 6) the *Philosophical Transactions*. The compilers also include an "Appendix," which covers texts more tangentially related to medicine, such as literary descriptions and alchemical/chymical discussions. EMEMT thus gives a fairly comprehensive view of the printed output by English medical writers in the early modern period. As emphasized in the introduction to the book (pp. 2-6), this setup particularly allows for the study of the intersection between developments in medical thinking and procedure and language use, to see whether, for example, the gradual transition from medieval, scholastic approaches to empirical methods is reflected in the language of the texts. The variety and stratification of texts also enables investigations of the connection between language and type of text, type of author (e.g. learned authors vs. authors with little or no formal training), or type of intended audience (e.g. learned physicians vs. surgeons).

As underscored by Irma Taavitsainen and Jukka Tyrkkö in their presentation of the categorization of the texts, it is important to recognize that EMEMT represents the world of medical texts in *print*; handwritten manuscript texts were not included in the corpus. Although this is a reasonable limitation, I wish the compilers had compensated for this exclusion by adding a more in-depth, focused discussion of our current knowledge of the copying and transmission of medical manuscript texts. Recent research has shown that some medical practitioners engaged in intense copying of texts. The Elizabethan astrological physician Simon Forman (1552-1611), for example, has left us numerous hand-written documents, attesting to his wide reading and excerpting from manuscripts as well as printed texts (Kassell 2005). Drawing on this research would have helped users who are new to the field of medical texts to gauge the extent and nature of manual copying vis-à-vis printing, and to get a sense of the genres and discourse forms that are available only in manuscript or that overlap with print. Such an exploration would probably also have led to a reconsideration or at least problematization of the compilers' claim that "printed texts arguably carried higher prestige" (p. 59) than manuscript texts.

The division of texts into larger categories has been a perennial problem for corpus compilers, and there is frequent debate about what features to take into consideration in assigning a text to a category. The classification of texts in EMEMT seems to have been particularly tricky

because of the many different forms and genres covered under the domain of medical writing. To provide categories, the compilers seem to have been forced to resort to a very broad classification, which appears to be primarily based on content. For example, texts included within the category General treatises and textbooks "range from learned and authoritative textbooks of medicine to all-in-one books providing access to basic theories of medicine and their applications" (p. 66). However, the category *Philosophical Transactions*, which represents extracts from the journal published by the Royal Society in the seventeenth century and beyond, is based on "publication format" (p. 127). This puts the categories on uneven footing since the classification is not based on uniform principles. Furthermore, within these larger units, there are several sub-categories or "genres": the *Philosophical Transactions* consists of book reviews as well as experimental reports; and recipes occur in several of the larger categories, the assignment of category being based on whether the text treats multiple substances and/or diseases, or one substance alone (p. 103). Although the compilers' decision to provide very general categories is understandable, it remains unclear how internally consistent the categories are and what the potential impact on linguistic studies may be of having, for example, reviews and experimental reports treated as one category. Depending on research question and aim, users may thus want to explore other, narrower classifications of the texts (see the discussion on software below).

The book that accompanies EMEMT is divided into four major sections: 1) Background; 2) Corpus Description; 3) Studies; and 4) Technical Aspects. While corpora are usually released with a manual and, in some cases, a brief description of the context of the texts, the EMEMT package clearly sets a new standard for future publications of corpora and text collections. With an enlightening introduction to medical practice in early modern England, to discourse forms and genres, and to the compilation and technical aspects of the corpus, EMEMT is made much more accessible to users who are not familiar with the domain of medicine (although some texts are still challenging owing to their technical nature). As historical linguists, if we are to understand the language of these texts, we need to understand the textual, communicative, and social context: this book provides us with the tools to do that.

The book is overall highly informative and readable. Especially readable are the discussions of the different text categories in Section 2 (written by various constellations of compilers). Among the chapters in the other sections, the meticulous study by Belén Méndez-Naya and Päivi Pahta in Section 3 (Studies) is particularly impressive in terms of clarity and scope. Investigating the shifting paradigms of intensifiers (such as *full*, *well*, *right*, and *very*), they demonstrate convincingly that a focused, domain-specific corpus such as EMEMT can "help qualify the general findings of studies drawn from general-purpose, multi-genre corpora" (p. 213).

There are aspects of the book's description and framing, however, where I would have wanted to see more specificity or clarity. This especially concerns the concept of "science" and its overlap with the concept of "medicine." The authors sometimes refer to "scientific" writing/discourse in relation to their corpus studies (e.g. pp. 37, 51, 212), and at other times "medical" writing (e.g. pp. 52, 193). Medicine is undoubtedly part of the complex of practices that can be considered early modern science, but it is unclear to what extent corpus results based exclusively on medical texts can be extrapolated to scientific texts more generally since there were many other manifestations of "science" in the period. Greater clarity about how the two areas overlap as well as differ would thus have been welcome.

The technical aspects of EMEMT (including the software, the presentation of texts, and other features) reveal quite a few innovative features. These technical aspects are described in Section 4 of the book, which includes a meticulous, illustrated manual (written by Jukka Tyrkkö, Raymond Hickey, and Ville Marttila). With the help of this manual, I had no problems installing the EMEMT software on two different PCs (both running Windows 7). The software (created by Raymond Hickey in consultation with the compilers) enables various searches and other manipulation of the text files (provided in extended ASCII format). In the search program, the texts are presented in a tree structure, which has seven branches in accordance with the seven text categories of EMEMT. The user can re-classify the material into whatever grouping the user may want by changing the tree format to a list format. As I suggested earlier, such a re-classification may be useful since the main categories provided in EMEMT are quite broad. Furthermore, as is shown in the studies in the book and in the compilers'

previous work, a great deal of diachronic variation occurs in the language of medical texts in the early modern period. Unfortunately, the EMEMT does not come with a set periodization, which means that users will have to construct their own periods. Dividing the texts into periods seems only to be possible by clicking appropriate texts in the list format and running separate searches on the separate sets of texts.

I was able to replicate the studies carried out in the manual and to carry out similar searches without problems.[1] More advanced search options and a host of other useful features of the program are outlined in a more detailed manual that accompanies the software. Some procedures produced persistent error messages the first time around, but after I closed down the program and conducted the procedures again, they worked smoothly.

Two features of the software deserve special mention, as they point to interesting developments in the presentation of corpus texts. Each text in the corpus is linked with a description of the text. This description includes author information, a brief overview of the content of the text, the library reference to the specific version presented in the corpus, hyperlinks to the *Oxford Dictionary of National Biography* and *Early English Books Online* (if appropriate), and other information. This allows the user to become familiar with the history and context of the text very easily. Even more impressive is the inclusion of images taken from the original books, such as front pages and illustrations. The illustrations are particularly useful as they allow a comparison between text and image, enabling studies of the multi-modal character of the texts.

Finally, in addition to the original texts presented in their original spelling, the EMEMT package includes what is called a "standardized" or "normalized" version of the corpus. Any English historical text collection from before at least the eighteenth century will present challenges for users because of the spelling variation that was common before the establishment of a clear standard in English: the user will have to make sure that he/she has found all variant spellings of a word in order

[1] Some of the searches that are illustrated in the manual must have been carried out on a pre-publication version of the corpus since dates of texts are not the same in the current corpus and in the illustrations. For example, Image 17 (p. 234) includes "1550 Langton. Introduction into physicke," while the corpus now has "1545 Langton. Introduction into physicke." Some of the word counts also differ, although the number of search hits match.

to avoid inaccurate analysis and faulty statistics. In collaboration with researchers from Lancaster University, the compilers of EMEMT have produced a version where the spelling variation has been substantially reduced. This has been accomplished with the help of VARD (Variant Detector), a program that identifies and automatically modernizes[2] the spelling of a word. It is claimed that this modernization "[makes] analysis easier and more accurate" (p. 284) and that we can "gain more precise results" (p. 289) by using the modernized texts. Although I would agree that it may help to make searches easier, since the user would perhaps not have to find a large number of variant spellings, I do think the claims of more accuracy and precision require modification. Modernization does not automatically provide more accurate or precise results: whether the user prefers the non-modernized or modernized texts for searches, the user would have to check for possible alternate spellings. About 73% of the variants detected by the program were modernized; the remaining variants did not reach a particular threshold where a modernized form could be assigned with confidence. Spelling variation may thus still occur, although the number of alternate spellings clearly depends on the nature of the search word. Furthermore, as with output from the non-modernized texts, careful post-processing will be needed even if the modernized texts are used, as the automatic modernization has clearly resulted in inaccuracies.[3] For example, it was decided that all verbal –th endings would be modernized to –s (with the exception of *hath* and *doth*, which are special cases; p. 286). However, this has led to a number of plural –th forms being modernized to –s. Among other instances, four examples of plural *sayth* (as in "Olde auncyent doctours of Physycke sayth…") in Andrew Boorde's 1542 *Dyetary of helth* have been turned into *says*. Consequently, the increased "precision" or "accuracy" of the procedure cannot be taken for granted, but must be evaluated carefully.

Overall, the EMEMT package (including the corpus, book, and software) is an important contribution to currently available corpora that cover aspects of the history of English, and to corpus compilation

[2] The authors of the chapter that describes the procedure (Anu Lehto, Alistair Baron, Maura Ratia, and Paul Rayson) refer to the process as standardization or normalization, but it is more accurately described as modernization, as the standard used is present-day spelling.

[3] There is no mention that the modernization was spot-checked after completion.

methodology and practice. Several features of the corpus (e.g. the information on texts, and presentation of images) and the book (e.g. the description of socio-historical context) set new standards for publications of corpora, and the material itself will undoubtedly give us new insights into the special language of medicine as well as the development of English in general.

Peter J. Grund
University of Kansas

References

Kassell, Lauren. 2005. *Medicine and Magic in Elizabethan London: Simon Forman: Astrologer, Alchemist, and Physician*. Oxford: Clarendon Press.

Taavitsainen, Irma, Päivi Pahta, and Martti Mäkinen. 2005. *Middle English Medical Texts*. CD-ROM. Amsterdam/Philadelphia: John Benjamins.