Irma Taavitsainen (University of Helsinki), Päivi Pahta (University of Tampere), Turo Hiltunen,
Anu Lehto, Ville Marttila, Maura Ratia, Carla Suhr (University of Helsinki), Jukka Tyrkkö (University of Tampere)

Introducing the Corpus of *Late Modern English Medical Texts 1700–1800*

Late Modern English Medical Texts 1700–1800 (LMEMT) is the last component of the three-part *Corpus of Early English Medical Writing 1375–1800*, compiled by the Scientific Thought-styles team. Together the three corpora offer material for larger diachronic study. LMEMT will facilitate further diachronic studies of linguistic processes in the special language of medicine.

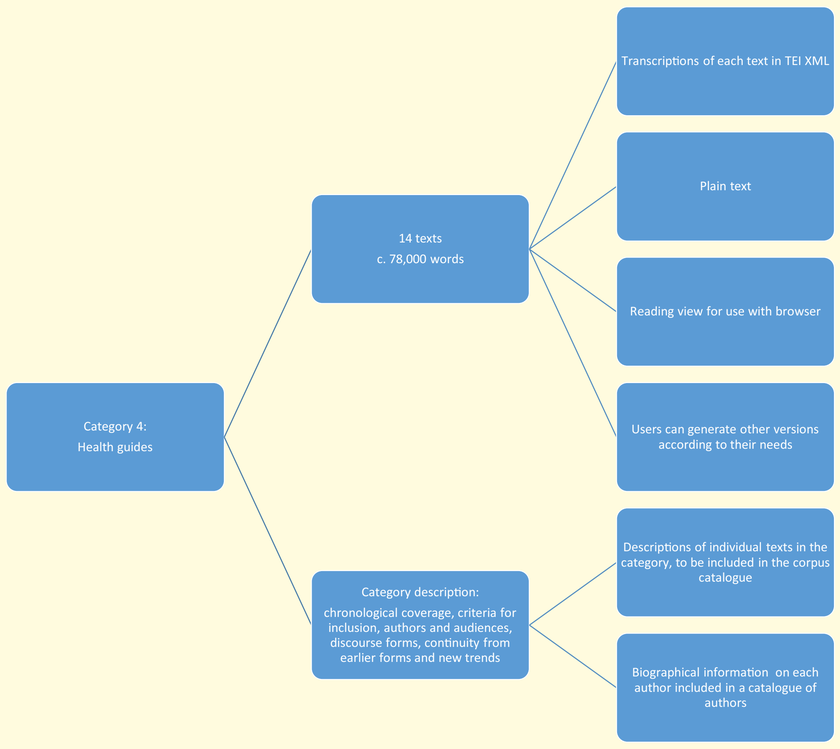

Figure 1 illustrates the contents of one category, health guides.

WHY MEDICINE?

- Medicine was the spearhead of the scientific register in the history of English
- Much of its long diachrony is still understudied and remains only partially understood
- The diachronic development of conventions in scientific and medical writing is an important area of research: understanding present-day practices of scientific writing is possible only if we know about past conventions and can relate them to the changing patterns of scientific thinking.

WHY THE 18TH CENTURY?

- The period 1700–1800 marked a transition from the thought-styles of the earlier periods to more modern approaches to medicine. The field of medical writing became increasingly complex in the 18th century.
  - The idea of public health emerged, and the numerous hospitals established in the 18th century became centres for developing new medical knowledge
  - Quantification and statistical methods began to be used in medicine
  - Obstetrics emerged as a sub-discipline of medicine
- Large systematically collected materials have not been available to linguists
- Provides continuation for *Middle English Medical Texts 1375–1500* (MEMT) and *Early Modern English Medical Texts 1500–1700* (EMEMT)

As the previous corpora, LMEMT will provide tools to probe into the socio-historical and cultural contexts of texts. The inclusion of a wide range of texts, both learned and popular, makes it possible to address entirely new research questions combining insights from separate fields of study. The corpus enables interdisciplinary research at the interface of corpus linguistics, philology, history of science, and book history.

CONTENTS AND STRUCTURE

- LMEMT contains c. two million words
- The corpus reflects an inclusive view of medicine so that it covers both elite and household practices
- A full range of texts are included from academic treatises to writings targeted at heterogeneous lay people, selected in collaboration with medical historians
- The main source of corpus texts is the online repository *Eighteenth Century Collection Online* (ECCO). Through institutional collaboration with the ECCO Text Creation Partnership (TCP) based in Michigan, we received some of the texts in XML format. In addition, a number of texts have been obtained by agreements with various repositories, and they have been keyed in.
- The catalogue contains descriptions of individual texts as well as short biographical data of authors. All the descriptions are collected into a searchable catalogue. The information in the catalogue can be used to create new classifications.
- The categorising is based on the topic of the text, or the area of medicine it represents.
- Texts within a category vary according to, for example, target audience or the author's educational background. Corpus users can adjust our categorisation to better suit their individual research questions.

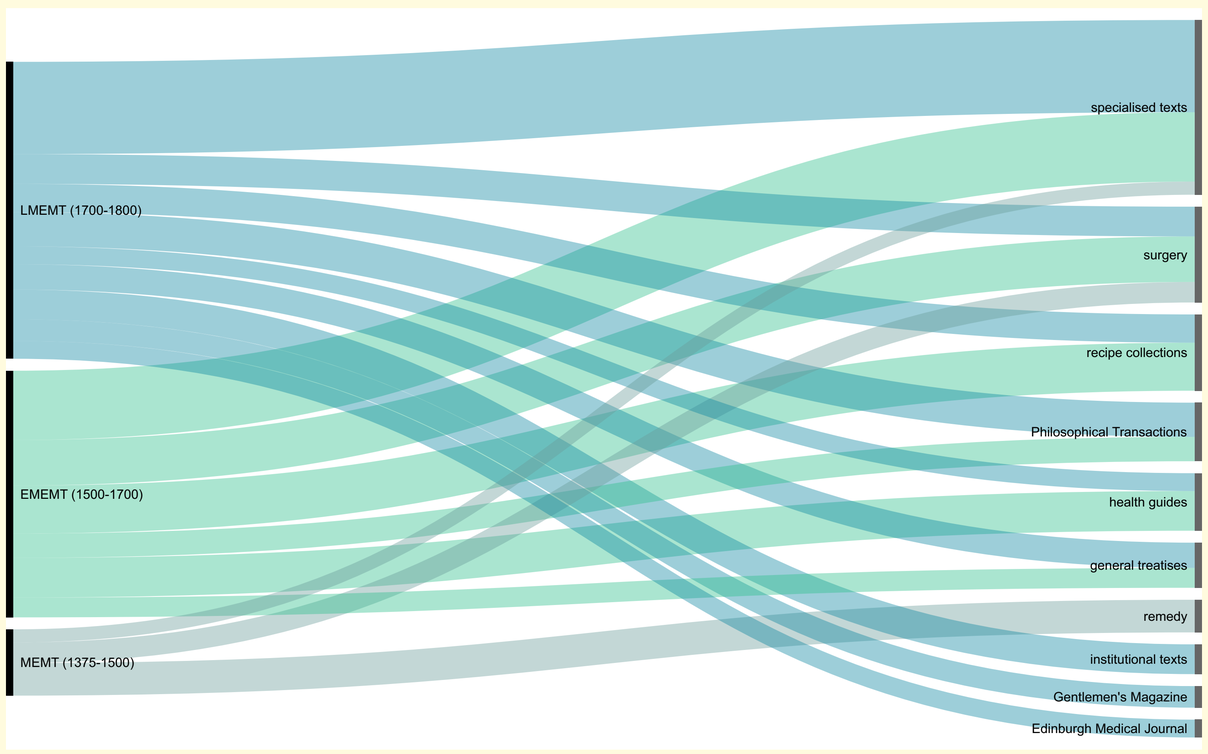

Figure 2 represents schematically the relationship between the categories used in the three corpora (MEMT, EMEMT and LMEMT) as well as their relative sizes.

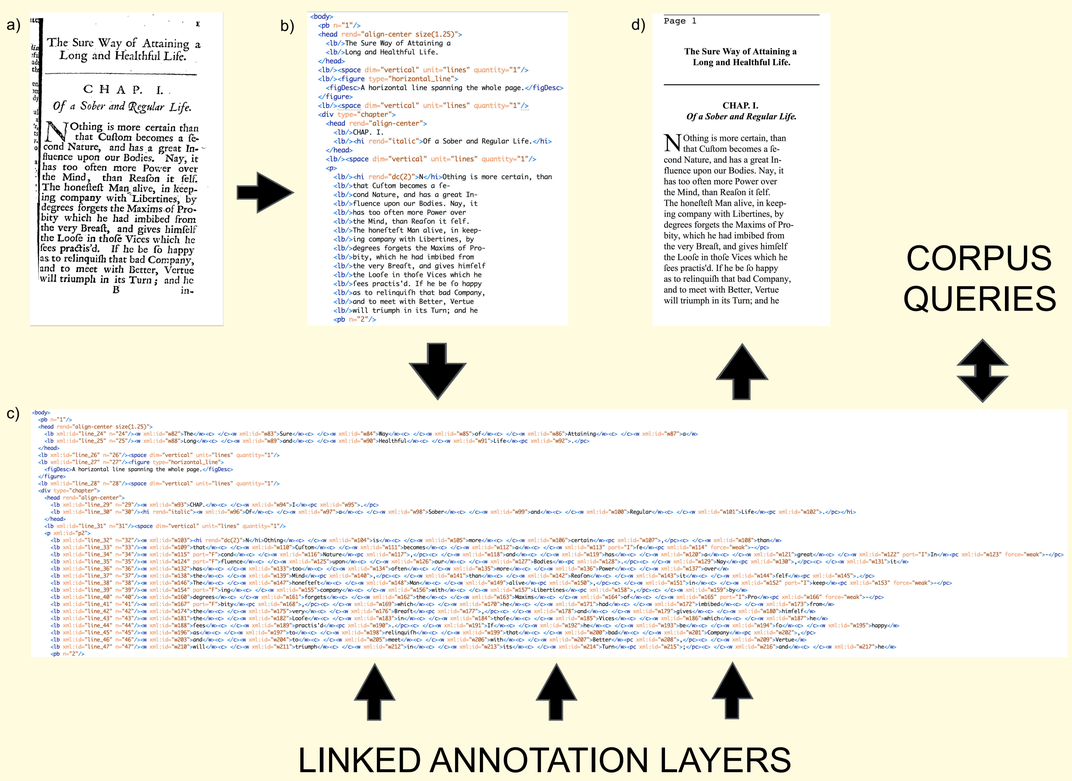

Figure 3 shows the different phases in LMEMT compilation and use: transcription from ECCO facsimiles (a) into descriptively annotated XML (b); automatic and semi-automatic analytical annotation, resulting in finished corpus data file (c); rendering of human-readable presentations of the corpus data (d) using style-sheets, selective extraction of data, and the linking of external analytical annotation to the corpus.

TECHNICAL DETAILS

- Corpus data files are encoded in XML following the *Guidelines for Electronic Text Encoding and Interchange* developed by the TEI Consortium
- Corpus texts annotated with:
  - textual structure (chapters, paragraphs, headings, notes, etc.)
  - visual layout (line, column and page changes, footnotes, indentation, illustrations, etc.)
  - typography (typeface, type size, etc.)
  - individual word-units with unique and persistent identifiers
- Extensive metadata formally encoded into corpus file headers:
  - bibliographic details of the source document
  - description of the significance, characteristics and content of each text, with keywords
  - documentation of transcription and annotation principles and practices
  - detailed biographical data about the author and identification of other individuals involved in the production of the text (translator, publisher, printer, etc.)
- Both metadata and inline annotation can be used to search and display the texts using suitable tools
- Allows for the linking of new annotation layers – such as part-of-speech tagging, syntactic parsing or semantic markup – to the corpus using stand-off markup
- Includes also stylesheets and linking files for conveniently browsing and reading the corpus files, and plain-text versions for use with legacy corpus software