The Lampeter Corpus of Early Modern English Tracts

– Am I really as dull as a tract, my dear? –
(G. Meredith in** Diana, **p. xviii)[1]

Rainer Siemund and Claudia Claridge
Chemnitz University

The *Lampeter Corpus of Early Modern English Tracts,* compiled at the Chemnitz University of Technology, has recently been completed and will shortly be available through *ICAME* and the *Oxford Text Archive*. The following article gives a brief outline of the corpus set-up and its coding conventions, and places it within its historical and corpus-linguistic background. For a more detailed account of the corpus, see the manual of information, *'Life is ruled and governed by opinion': The Lampeter Corpus of Early Modern English Tracts* (Schmied, Claridge and Siemund forthcoming), which will be published by *ICAME* this autumn.

1 Lampeter and the corpus

The name *Lampeter Corpus* derives from the place where the corpus texts were gathered, namely the Founders' Library at the University of Wales, Lampeter. Founded in the 1820s as St. David's University College for religious education, the institution owned a comprehensive library, not only of religious interest, from its earliest days. The majority of the Tract Collection, from which the corpus material is taken, was a bequest of the Bowdler family, the most prominent member of the family being the editor Thomas Bowdler (1754–1825), whose attempts to clear Shakespeare's works of indecent passages earned him a dubious reputation – one which has lasted down to the present day. Luckily the Tract Collection was not bowdlerised at all, and it presents a cross-section of the literature published in the form of tracts or pamphlets in the 17th and 18th centuries (cf Harris and James 1974). The full catalogue of the Founders' Library is accessible on the World Wide Web through the Online Public Access Catalogue of Lampeter University.

2 *The corpus and the texts*

Origin of the corpus

The Lampeter project was launched in 1991 at Bayreuth University by Josef Schmied, the current project leader and director of the *REAL* (Research into English and Applied Linguistics) *Centre* at Chemnitz University. The project has been funded since 1994 by the *Deutsche Forschungsgesellschaft* (*DFG*), the German Academic Research Association. A number of changes have been made to the original set-up of the corpus over the years, but the focus on non-literary prose literature has remained the same, as has the sampling period.[2] The latter covers the century between 1640 and 1740, spanning three generations with respect to language change and reaching, in historical terms, from the outbreak of the Civil War to the first flickers of the Industrial Revolution a hundred years later. From a linguistic point of view, it is especially the standardisation process and the increasing use of English as an all-purpose language that make the period covered by the *Lampeter Corpus* an interesting one. The word *corpus* implies that the *Lampeter Corpus* is not a random collection of texts but a 'principled' (Johansson 1991:3) effort to mirror the range of publications at the time.

Pamphlets and tracts

According to the *OED*, a pamphlet is 'a small treatise occupying fewer pages or sheets than would make a book', and is usually printed and issued as a separate work. In the 17th century, the term was generally 'used irrespective of subject', could include 'issues of single plays, romances, poems, novelettes, newspapers, news-letters, and other periodicals', and occasionally referred to as chap (= cheap) books. The selection of the corpus material is determined by this fairly general understanding of the term. In order to distinguish our collection of texts from today's rather narrow understanding of 'pamphlet' as a publication chiefly political in tone, however, we decided to use the more neutral term 'tracts' for the corpus title. At the beginning of our sampling period, a tract was considered to be a 'book or written work treating of some particular topic, a treatise, a written or printed discourse or dissertation'; in the course of the 18th century, the sense of the word developed into a 'short pamphlet on some religious, political, or other topic, suitable for distribution or for purposes of propaganda' (*OED*). As the definitions of 'pamphlet' and 'tract' show, the *Lampeter Corpus* cannot be subsumed under a single text type or genre. Texts vary in format, length, argument structure, function and topic, and to arrange them under a single genre category would be as inadequate as to attribute a single text type to books. Tracts, like books, are a production type or publication form rather than a genre or text type.

Corpus structure

The basic structure of the corpus is a division into decades, ranging from the 1640s to the 1730s. Within each decade, the texts mirror the variety of topics treated in contemporary publications, covering the domains *religion, science, economy and trade, law, politics* and *miscellaneous*. Each of these six domains accommodates two texts per decade, yielding a total for the sampling period of 120 texts written by 120 different authors. With a few exceptions due to the material available, the short tracts in the corpus consist of originally unbound, separate publications ranging in length between 3,000 and 20,000 words, amounting to some 1.1 million words for the whole corpus. As the varying lengths of the texts indicates, the corpus is not made up of samples cut off at a certain point convenient for statistical analysis but consists of complete texts from title page to obligatory *FINIS*, including addresses to the reader, dedications, prefaces, and texts proper, as well as postscripts and appendices. The material therefore allows for the text-linguistic and stylistic analysis of all the documents or subdivisions of them. Corpus users who require only samples will find markers delimiting 3,000-word segments within the running text.

Even though anonymous publications were a sign of the times, we have kept their number as low as possible so as to be able to provide the necessary background material on authors for variation studies. The framework for additional information of different kinds is supplied by text headers that are attached to each of the documents proper (see below). Most of the tracts were printed in London due to the organisation of printing in early modern Britain and the licensing restrictions of the press, especially during the first decades of the sampling period. This does not mean, however, that the topics covered were only of metropolitan interest or that the majority of authors were London-based. An increasingly sophisticated retailing system, first by post, later also by local booksellers, ensured that the tracts could reach even the remotest parts of the British mainland. Readers were able to participate actively or passively in the ongoing debates of the time. To avoid any idiosyncrasies of individual authors, each of them appears only once in the corpus. We decided to exclude the major literary figures of the time, since their use of language can be studied elsewhere. To ensure the originality of language, only the first issues of tracts were included, apart from a few later copies that were substantially revised or enlarged by their original authors for re-publication. All of them are contemporary, however, and were keyed in directly from the 17th- and 18th-century documents without intermediary modern editions of the texts: the corpus thus retains the original orthography, punctuation and word divisions. The majority of the tracts appear in the *Lampeter Corpus* for the first time since their original publication. In addition, the corpus gives background information on the texts that is not present in the originals; this is done by means of *TEI* headers.

TEI/SGML encoding

The *Lampeter Corpus* uses the coding scheme suggested by the *Text Encoding Initiative (TEI)*, an organisation providing guidelines for the interchange of electronic texts. It consists of two parts. One is a system of corpus and text headers preceding the texts proper, in which background information on texts and markup is stored in order to keep the running text largely free from extralinguistic information. The second part involves the use of the *Standard Generalized Markup Language (SGML)*, so that information on structure and layout of the texts is retained even after the corpus has been reduced to pure *ASCII* format. *SGML* is an internationally recognised standard for information interchange (ISO 8879) and thus ensures a far-reaching compatibility of the texts. It can also serve as the basis for Internet versions of texts, since the *Hypertext Markup Language (HTML)* is a version of *SGML*.

The text headers of the *Lampeter Corpus* also make available additional material from different areas. To begin with, they contain information such as the authors' names (it proved possible to trace the authors for most of the anonymous publications), age, sex, place of residence, education, and social status. Headers also point to printers and publishers (whose influence on the use of language or contents has yet to be determined), and also print place and date (which are not always present on the title page). They record the format of the text (folio, quarto or octavo), give references to bibliographical works such as *Wing's Short*

Title Catalogue or to shelf and catalogue numbers in the Lampeter library, list the total number of words in a file, and enumerate the kinds of composite parts of a text, such as preface, dedication or introduction. The original plan to incorporate a three-dimensional text typology into the headers has been abandoned. (For a detailed discussion of text types in the *Lampeter Corpus*, see Schmied and Claridge 1997.) Only the self-description of the texts has been retained to provide a starting point for genre analysis from an early modern perspective.

The system of text headers is complemented by *SGML* marked-up text. At the outset of the project, the obvious choice for textual markup seemed to be the system used by the *International Corpus of English (ICE)*, especially since the *REAL Centre* is also compiling the East African part of *ICE* and the markup system has been especially designed for linguistic applications. In the course of time, however, an increasing number of additions and alterations had to be made to account for the different character of a historical corpus to that of a corpus covering national varieties, so that in the end the texts were altogether converted into SGML. By using a system of begin and end tags (<...> and </...> respectively) on all text levels, *SGML* coding ensures that all the structural and layout characteristics of the *Lampeter* texts can be recaptured.<*text*>, for example, indicates the beginning of a corpus text, <*p*> the beginning of a paragraph, <*hi rend=italics*> the beginning of a highlighted element printed in italics. *TEI/SGML* aware software is able to either suppress or render visible the marked-up information – depending on the specific research questions. It also allows searches in selected parts of the corpus, for example of texts printed outside London, written by under-30-year-olds or by members of the lower gentry. *SGML* browsers, then, can restore the original layout to the computer screen and to print.[3]

3 *The texts and historical context*

It was important for us that the corpus format be easily adaptable for different kinds of use, as the texts are intended as a tool not only for linguistic but also for literary and historical research. The sampling period, 1640 to 1740, lies at a crucial point in the formation of modern society, and the classification of the texts according to subject matter is an attempt to show developments in *religion, science, economy and trade, law,* and *politics*. The tracts contained in the *Lampeter Corpus* played an important part in the shaping of public opinion and comment on a host of events, some of which are outlined below.

Religious life, for example, saw the clashes of Protestantism with Catholicism – and within Protestantism itself – that determined the course of politics throughout the 17th and early 18th centuries. The fear of ‘Popery’ shaped both religious and political rhetoric at a time when English Protestants saw England as one of the last strongholds against ‘dissenters’ from within and without. Religious groups of all kinds developed who sought and found confrontation with traditional religious doctrines. The rise of Protestantism goes hand in hand with the rise of early modern science. The trial against Galileo in 1633, shortly before the start of our sampling period, was one of the last major drawbacks scientific progress had to suffer. As a consequence of a social climate largely free from religious taboos during the second half of the 17th and throughout the 18th century, scientific societies were founded, journals established, and mathematical techniques invented and applied to physical problems. The Copernican system won widespread acceptance, objects for the observation of very small and very distant objects became widely used, and a scientific discourse began to develop. Scientific inventions and the elaboration of modern technologies played an important part in the expansion of British trade. Nautical instruments invented in the 17th century and increasing knowledge about celestial phenomena facilitated navigation and reduced the risks at sea to a more predictable measure. The British economy, and with it newly established trading corporations like the East India Company, thrived on the establishment of British dominions in foreign parts of the world. These new enterprises required a different financial and legal background than that of traditional and less capital-intensive businesses within the mother country. By providing the financial and the legal framework necessary for the ventures undertaken by joint-stock associations, metropolitan bankers and lawyers increasingly became the support on which the economy rested. The central legal question between 1640 and 1740, however, concerned the relation of Common Law to the Royal Prerogative. The absolutist view generally taken by the Stuart dynasty and religious unrest throughout the country led the country into civil war and remained a powder keg until the question of succession to the English throne was put to rest with the beginning of Hanoverian rule in 1714. A by-product of the ensuing shift of power from Court to Commons was the rise of party politics and the lapse of the licensing system of the press in 1695. This, in turn, led to a hitherto unknown output of the printing presses, and the controversial issues debated in the tracts of the *Lampeter Corpus* testify to this new liberty.[4]

4 *The Historical context and historical corpus linguistics*

Corpus linguistics today can look back on three generations of corpora. First-generation sample corpora such as *Brown* and *LOB*, primarily compiled as accessible tools for non-commercial linguistic study, were succeeded by multi-million word corpora that were often supported by publishing houses for lexicographic research. Yet in the early 1980s, writes Sinclair (1991:9), 'as the multi-million-word corpus became available for study, it became clear that the whole idea of a corpus of finite size was flawed. Any corpus is such a tiny sample of a language in use that there can be little finality in the statistics.' This knowledge led to the idea of creating the third-generation monitor corpora that make use of the electronic print media mailing their daily production onto mainframe computers, thus making possible several-hundred-million-word collections of text.

Historical corpus linguistics and the corpora it produced still belong to the first generation, and everyone who has ever keyed in texts manually from old manuscripts or early printed documents knows the reason why. Mass production of computer-readable Old English, Middle English or Early Modern English texts, not to mention speech transcriptions, is highly unlikely. However, first-generation corpora do have their merits. The first one is that they are readily available for use on a PC, rendering them an easily accessible and profitable teaching and study aid. Secondly, a corpus should be different from a mere collection of texts or an archive. In contrast to many second-generation and third-generation corpora, historical corpora are compiled along 'principled' lines, and even though they cannot always offer the final word on the phenomena under investigation, they provide a basis for further analysis. According to Sinclair (1991:100), 'the received wisdom of corpus linguistics is that fairly small corpora, of one million words or even fewer, are adequate for grammatical purposes, since the frequency of occurrence of so-called grammatical or function words is quite high.'

An alternative to multi-million word corpora can be provided by a network of first-generation corpora, however, and ideally, hypotheses formed on the basis of one corpus can be verified or falsified through the analysis of related ones. The reports on developments in historical linguistics in this volume of the *ICAME Journal* show that a lot of work has been done, especially in the field of Early Modern English. Each corpus provides a piece in a jigsaw puzzle and an increasing number of pieces contributes to an increasingly complete picture of language in the period. The *Lampeter Corpus* complements other projects such as the early modern English part of the *Helsinki Corpus* (1640–1710), the *Archer Corpus* (1650 to the present day), the *Century of Prose Corpus* (1675–1775), the *Zurich English Newspaper Corpus* (mid-1660s to the beginning of the 20th century), the *Corpus of Irish English* (12th to 20th century), the *Helsinki Corpora of Older Scots* (1450–1700) and *Early American English* (early 17th to early 18th centuries), the *Corpus of Early English Correspondence* (1420–1680), the *Corpus of Dialogues* (1550–1750), several computer-readable editions of early modern authors, and a number of dictionaries – some of them still under development. Research on Early Modern English has turned into a rapidly growing enterprise and is on its way towards having a multi-million word corpus available for linguistic study. As the number of recent publications based on Early Modern English corpora shows, the unjustified neglect of the period in the past has been largely remedied.

The *Lampeter Corpus* and an accompanying manual, as mentioned above, will shortly be available from *ICAME* and the *Oxford Text Archive*. For questions concerning the project or related issues, please visit us at *http://www.tu-chemnitz.de/~ehe/real.htm* or contact us via

Prof. Dr. Josef Schmied
The REAL Centre
Chemnitz University of Technology
Department of English Language and Linguistics
D– 09107 Chemnitz

FINIS.

Notes

[1] *OED*, 2nd ed. on CD-Rom.

[2] For the original corpus set-up cf Schmied 1994.

[3] Several introductions to *SGML* are at large. A manageable one is that written by Lou Burnard and C.M. Sperberg-McQueen (1995). For further information on and around *SGML*/TEI, including software, see the Oxford Text Archive on the World Wide Web, http://sable.ox.ac.uk/ota

4 For introductory literature on some of the topics covered in the *Lampeter Corpus* as well as on printing and literacy, see Cressy (1980), Hall (1981), Cain and Robinson (1992), Raven (1992), Eisenstein (1993), Hill (1993), Babington (1995) or Sharpe (1997).

References

Babington, Anthony. 1995. *The rule of law in Britain. From the Roman occupation to the present day*. 3rd ed. Chichester: Barry Rose.

Burnard, Lou and C.M. Sperberg-McQueen. 1995. *TEI Lite: An introduction to Text Encoding for Interchange*. World Wide Web at *http://www-tei.uic.edu/orgs/tei/intros/teiu5.tei*

Cain, Thomas Grant Steven and Ken Robinson (eds). 1992. *Into another mould. Change and continuity in English culture 1625–1700*. London: Routledge.

Cressy, David. 1980. *Literacy and the social order*. Cambridge: Cambridge University Press.

Eisenstein, Elizabeth L. 1993. *The printing revolution in early modern Europe*. Cambridge: Cambridge University Press.

Hall, A. Rupert. 1981. *From Galileo to Newton*. New York: Dover.

Harris, L.J. and B.L. James. 1974. The tract collection at Saint David's University College, Lampeter. *Trivium* 9, 100–109.

Hill, Christopher. 1993. *The century of revolution 1603–1714*. London: Routledge.

Johansson, Stig. 1991. Times change, and so do corpora. *English corpus linguistics. Studies in honour of Jan Svartvik*. Karin Aijmer and Bengt Altenberg (eds). London & New York: Longman, 305–314.

Kytö, Merja, Matti Rissanen and Susan Wright (eds). 1994. *Corpora across the centuries. Proceedings of the First International Colloquium on English Diachronic Corpora, St. Catharine's College Cambridge, 25–27 March 1993*. Amsterdam & Atlanta: Rodopi.

Oxford English Dictionary on CD-ROM. 1992. 2nd ed. Oxford: Oxford University Press.

Raven, James. 1992. *Judging new wealth: Popular publishing and responses to commerce in England, 1750–1800*. New York: Clarendon Press of Oxford University Press.

Schmied, Josef. 1994. The Lampeter Corpus of Early Modern English Tracts. *Corpora across the centuries*. Merja Kytö, Matti Rissanen and Susan Wright (eds). Amsterdam & Atlanta: Rodopi, 81–89.

Schmied, Josef and Claudia Claridge. 1997. Classifying text- or genre-variation in the Lampeter Corpus of Early Modern English texts. *Proceedings of the 16th ICAME conference. Diachronic volume*. Raymond Hickey, Merja Kytö and Matti Rissanen (eds). Amsterdam: Rodopi.

Schmied, Josef, Claudia Claridge and Rainer Siemund. Forthcoming. *'Life is ruled and governed by opinion': The Lampeter Corpus of Early Modern English Tracts*. Bergen: Norwegian Computing Centre for the Humanities.

Sharpe, J.A. 1987. *Early modern England. A social history 1550–1760*. London: Edward Arnold.

Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.