

English historical corpora: report on developments in 1998

Merja Kytö and Matti Rissanen
Uppsala University and University of Helsinki

The present report will supplement those included in *ICAME Journal* 19 (1995), 20 (1996), 21 (1997), and 22 (1998) (for further information on corpus-work done within the diachronic approach, see, eg *Corpora Across the Centuries: Proceedings of the First International Colloquium on English Diachronic Corpora*, eds Merja Kytö, Matti Rissanen and Susan Wright, Amsterdam and Atlanta, GA: Rodopi, 1994, and *Tracing the Trail of Time. Proceedings from the Second Diachronic Corpora Workshop*, eds Raymond Hickey, Merja Kytö, Ian Lancashire and Matti Rissanen, Rodopi, 1997).

English historical corpora and the work done on them were the topic of the recent workshop organized in Newcastle (Northern Ireland) prior to the 19th ICAME conference in May 1998.

We are indebted to the scholars working on corpus studies for sending us their contributions for this report.

Matti Rissanen: matti.rissanen@helsinki.fi
Merja Kytö: merja.kyto@engelska.uu.se

NEW CORPUS PROJECTS

The York-Helsinki Parsed Corpus of Old English Poetry

This corpus project will produce an annotated version of the poetic texts of the Old English section of the Helsinki Corpus. The texts will be part-of-speech tagged and syntactically tagged and bracketed; the annotation scheme used is the same as that of the Penn-Helsinki Parsed Corpus of Middle English II. The

project is funded by a 30-month research grant from the Economic and Social Research Council (United Kingdom), and will be completed by December 2000.

Susan Pintzuk: sp20@york.ac.uk

The English language of the north-west in the late Modern English period

The recent history of the English language is a field neglected until recently but likely to become of more central interest with the publication of the 1776-1997 volume of the *Cambridge History*. Historians of English lack good collections of material in electronically-readable form from that period, apart from 'ARCHER' and Denison's 'Late ModE Prose' corpus (1994), and especially material written by people locally based and unused to writing for publication, which can provide particularly telling evidence. The John Rylands University Library of Manchester holds several collections of documents from the period, including some containing letters with personal material. The aim of the project is to select suitable documents, transcribe them, put them in machine-readable form with parallel original-spelling and normalised texts (necessary for systematic searching), and – subject to the appropriate permissions – publish them for scholarly purposes on CD-ROM or on the WWW. The texts will be useful to linguists and to social historians.

In the initial phase, letters written to Richard Orford in the Leghs of Lyme collection will be collected. In principle this is a project which could later be extended in scope, for instance to include other letters from the Legh, Bramley Davenport, and perhaps Dunham Massey collections, especially if the first collection of material is well used by scholars.

The project is directed by David Denison in collaboration with Linda van Bergen. With the aid of a bursary from the John Rylands Research Institute, Ms van Bergen is working for one day a week during 1998–99 to select, transcribe and annotate the Orford documents. Non-textual coding at present is limited to bare identification of each document's writer and date. By the end of summer 1999 it is estimated that some 120,000 words will have been transcribed and proof-read.

David Denison: d.denison@man.ac.uk
<http://www.art.man.ac.uk/english/staff/dd/>

PROGRESS OF EARLIER PROJECTS

The Brooklyn-Geneva-Amsterdam-Helsinki Parsed Corpus of Old English

The corpus project has produced a glossed, morphologically tagged, and syntactically tagged and bracketed version of prose texts of the Old English section of the Helsinki Corpus.

Two groups of scholars from three countries collaborated on the project. The first group included Ans van Kemenade, Willem Koopman, and Frank Beths (Amsterdam, the Netherlands), and was responsible for the morphological tagging of the corpus; the second group included Susan Pintzuk (York, England) and Eric Haeberli (Geneva, Switzerland), and was responsible for glossing and syntactic tagging and bracketing. Pintzuk's work was supported by a grant from the National Endowment for the Humanities (USA), an independent agency.

The morphological tagging of all of the prose texts in the Helsinki Corpus has been completed by the Amsterdam researchers, and this version of the corpus is available from Ans van Kemenade. 100,000 words of text have been syntactically annotated and glossed; and this version of the corpus, accompanied by a manual and search tools, will be available in August 1999 from Susan Pintzuk.

Susan Pintzuk:

sp20@york.ac.uk

Ans van Kemenade:

kemenade@let.vu.nl

Corpus of Early English Medical Writing 1375–1750

The Corpus of Early English Medical Writing 1375–1750 is under work in the English Department of the University of Helsinki by Irma Taavitsainen and Päivi Pahta. Its total size is ca one million words (March 1999); we aim at ca 1.5 million. The medieval part of the corpus extends from 1375 to 1550, and it is further divided into two periods, 1375–1475 and 1475–1550. We consider this level of precision adequate, as most texts lack exact dates and other precise background information. The medieval part of the corpus contains ca 535,000 words. Text selection pre 1475 is nearly completed, and the earliest part of the corpus is as comprehensive as possible, including all texts known to us and available to us. The underlying traditions of academic texts, surgical treatises and remedybooks are all represented. The corpus includes surgical and anatomical texts, special fields, encyclopaedias and compendia, remedybooks and recipes, and medical verse from the earliest period. Shorter texts are included in toto, and more comprehensive treatises are represented by extracts of at least

10,000 words. Work on the Early Modern part of the corpus has been started (for the overall plan, see *ICAME Journal* 21, 1997, pp 71–78). Text selection for the latter period has to balance various traditions and subgenres in the widening use of English for scientific purposes. At present the corpus is not available for public use.

Irma Taavitsainen: irma.taavitsainen@helsinki.fi
Päivi Pahta: paivi.pahta@helsinki.fi

ICAMET – Innsbruck Computer Archive of Machine-Readable English Texts

The MANUAL TO ICAMET has recently appeared. It describes the three parts of ICAMET: the corpus of Middle English prose (with 128 full-text data bases); the corpus of 254 English letters written from 1386 to 1688; and a (less important) collection of various other Middle English and Modern English texts ('VARIA'), mainly translations or normalised versions of Middle English texts. Moreover, the MANUAL gives evidence of the principles of compiling and encoding, and, of course, of the texts, sources and availability of the files. As author of the MANUAL I have tried to justify the way in which I have come to terms with the specific, generally underestimated problems posed by Middle English prose texts, and I would wish for some feedback – support, criticism, or suggestions – from fellow corpus linguists and medievalists.

Unfortunately, the distribution of the ICAMET texts on the INTERNET is not possible at present, due to copyright restrictions as imposed, in particular, by the Early English Text Society. However, some of the texts of the PROSE CORPUS and almost all the other text files, ie above all those of the letters, are available on diskette. The details are again given in the MANUAL. The MANUAL has just been published within our Innsbruck departmental publication series and can be ordered from your bookshop or from Universität Innsbruck, Institut für Anglistik, Innrain 52, 6020 Innsbruck, Fax +512-507-2882 (ISBN 3-85124-163-0; 120 pages; 190 ATS/29 DM/14,50 EU).

Present work within the Innsbruck ICAMET project is mainly concerned with evaluating corpus files and with completing the parameters of description soon to be published on the INTERNET. Apart from occasionally adding new texts to the PROSE corpus, we are also working on ways and methods of increasing the philological user-friendliness of the files, both by means of versions which are free of higher ASCII and by means of additional normalised versions. At present, the files are all stored in original MICROSOFT WORD for

DOS, with the historical signs, such as yogh and eth, to be produced by a special keyboard driver 'icamet'.

Manfred Markus:

manfred.markus@uibk.ac.at

Edinburgh Corpus of Older Scots (ECOS) / Older Scots Atlas Project

Work continues on compilation of ECOS. The focus remains on record and charter texts as the means of providing a matrix of localizable texts for the atlas project.

To date 532 texts have been lexico-grammatically tagged. The associated data-base of manuscript sources contains 440 entries. 120 disk-files containing transcribed texts await tagging. Geographical areas of coverage which require further attention are North-East and South-West Scotland, and effort is now being concentrated on gathering and processing texts with probable origins in these regions. For the present atlas pilot project dealing with the late 14th and 15th centuries, a reasonably representative sample of texts distributed geographically and chronologically is the objective. The aim is to achieve a good basic matrix in the coming year. From this, I hope to be able to produce exemplar results towards construction of the atlas.

Development of our software tools for managing, processing and analysing the corpus texts also continues. In particular, empirical testing of theoretical aspects of the work has been very much to the fore in the last year. And this will be carried on in tandem with augmentation of the corpus.

I aim to produce a monograph this year which will be a report on ECOS and its methodology in the context of the Edinburgh atlas projects (Early Middle English and Older Scots).

I am also collaborating with Dr Anneli Meurman-Solin, Department of English, University of Helsinki (compiler of the Helsinki Corpus of Older Scots (HCOS)) as part of that department's VARIENG research project on dialectology and regional variation. We plan three studies using the resources of HCOS and ECOS:

- (1) variation and variety in Older Scots texts and the value of a corpus-based methodology to study it;
- (2) the issues of linguistic tagging, in particular the information content of tags and the value of a largely theory-independent system of tagging for a wide range of studies;
- (3) the notion of a 'standard' Middle Scots and its linguistic validity.

Work in preparation

1999. Changing spaces. Linguistic relationships and the dialect continuum. In I. Taavitsainen, T. Nevalainen, P. Pahta and M. Rissanen (eds). *Language and Text: Selected Papers from the Second International Conference on Middle English* (Helsinki, 1997). Berlin: Mouton de Gruyter. [Paper originally delivered at the 10th International Conference on English Historical Linguistics, University of Manchester, 21–26 August 1998].
- Prolegomena to a Linguistic Atlas of Older Scots [Report of research funded by The Leverhulme Trust, 1994–1998].
- Things to do with lexico-grammatical tags. [For the 20th Annual Meeting of the International Computer Archive of Modern/Medieval English at Freiburg, 26–30 May, 1999].
- ‘*ye yhere and ye place for-wretyn*’ – spatio-temporal aspects of linguistic variation in Older Scots Texts. [For the 9th International Conference on Medieval and Renaissance Scottish Language and Literature, University of St Andrews, 7–11 August 1999].

Institute for Historical Dialectology,
School of Scottish Studies,
University of Edinburgh,
24 Buccleuch Place,
Edinburgh EH8 9LN,
Scotland

I.K. Williamson: i.k.williamson@ed.ac.uk

The Lampeter Corpus of Early Modern English Tracts

The news concerning the Lampeter Corpus is available at the internet address:
<http://tu-chemnitz.de/phil/english/real/lampeter/lamphome.htm>

Josef Schmied: josef.schmied@phil.tu-chemnitz.de
Claudia Claridge: claridge@mail.uni-greifswald.de
Rainer Siemund: rainer.siemund@phil.tu-chemnitz.de

Corpus of Early English Correspondence (CEEC)

The Corpus of Early English Correspondence has reached the final stage of compilation; no available original spelling editions of personal letters are likely to add much to the social coverage of Early Modern English, although the volume of the corpus could naturally be increased. At 2.7 million words covering the years 1417–1681, the corpus is now awaiting the clearance of copyright questions prior to publication. Due to lack of funding and the considerable number of texts included, this is likely to take some years yet, so in the meantime a sampler version of the corpus (CEECS) has been made available via ICAME.

More information at <http://www.helsinki.fi/doe/projects/ceec/> and *ICAME Journal* 23: 53–64.

Terttu Nevalainen:	terttu.nevalainen@helsinki.fi
Helena Raumolin-Brunberg:	helena.raumolin-brunberg@helsinki.fi
Arja Nurmi:	arja.nurmi@helsinki.fi
Minna Palander-Collin:	minna.palander-collin@helsinki.fi

CONCE – A Corpus of Nineteenth-century English underway

With the increasing interest in the study of late Modern English follows a need for corpora that enable an extension of the diachronic perspective afforded by the Helsinki Corpus. Such corpora should include core genres of the Helsinki Corpus to increase compatibility. Moreover, the recent interest in short-term language change (cf studies carried out on the LOB, FLOB, Brown, and Frown corpora) has created a need for a late Modern English corpus divided into relatively short subperiods.

A corpus of nineteenth-century English texts is being compiled at Uppsala and Tampere Universities. The corpus will contain specimens representative of a number of central text types such as fiction, science, history writing, drama, and correspondence. Moreover, trial proceedings and parliamentary debates are being added to the collection. The subperiodization will make it possible to study texts drawn from the beginning, middle and end of the century.

By August 1999, the corpus will comprise a good million words. Of valuable help in proofreading and other tasks has been Erik Smitterberg at the Department of English, Uppsala University, and the secretarial staff at the Department of English, Tampere University. At present the corpus is not available for public use, but all efforts will be made to offer access to the database in due course.

Merja Kytö: merja.kyto@engelska.uu.se
Juhani Rudanko: juhani.rudanko@uta.fi

HISTORICAL DICTIONARIES AND ATLASES

Dictionary of Old English project

The response of scholars to the free access to the Dictionary of Old English Corpus on the World-Wide Web (for a trial period from September 1998 through January 1999) has been encouraging. We are pleased that our latest research tool, distributed by the University of Michigan Press, has proved useful to so many of our colleagues. Information on site licenses (\$200 US per year per institution) is available from the project's webpage: <http://www.chass.utoronto.ca/oec/>

Now that the Dictionary of Old English Corpus on diskette and the Corpus searchable on the Web conform to the latest standard for text markup, we are focusing our attention on tagging the *Dictionary* itself. In our initial investigation into text transduction, we looked at various parsing packages. One of our goals in using a parser was not only to have a transduction but also to have as a by-product a grammar which could be used for the SGML (Standard Generalized Markup Language) DTD (Document Type Definition). However, the packages we looked at were inadequate for various reasons. Our visit to the Middle English Compendium Project at the University of Michigan was crucial in helping us determine how tagging should continue. As a result of this visit, we decided to craft an initial *perl* script to do the basic transduction, and then to tag parts of the text iteratively and develop the SGML DTD through the iterations. This process is currently underway with the initial conversion completed, and the refinement of the tagging continuing. We are first tagging our legacy data, which are the fascicles already published. Our basic model is to follow the internal, logical structure of the entry, with its twelve distinct fields, for the markup scheme. So far this process has been working well. As of 1 March 1999, two letters (*Æ* and *E*) of the six published fascicles of the *Dictionary* have been tagged. If we can continue at this pace, without too many technical problems, we should be in reasonable shape for CD-ROM production of the seventh letter, *F*, together with the six previously published letters, next year. In the meantime, entry-writing is continuing on the letters *F*, *G*, and *H*.

Antonette diPaolo Healey: healey@doe.utoronto.ca

For inquiries about the
Electronic Corpus: corpus@doe.utoronto.ca

Historical Thesaurus of English: Progress report March 1999

The Historical Thesaurus of English project lists the vocabulary of English from Old English to the present, arranged in detailed semantic categories along with dates of use based on the *Oxford English Dictionary*. The work will be published both as a book and electronically. Samples are available on our web-site <<http://www.arts.gla.ac.uk/EngLang/thesaur/homepage.htm>>.

Although at one point we had hoped to complete editing by the dawn of the Millennium (as popularly misdefined), we are not quite there yet. However, progress continues to be made despite the inevitable setbacks. 526,923 records out of an estimated total of 650,000 had been entered by the end of February 1999. Division III, Society, which is virtually complete, represents 44 per cent of this total, with Division I, the External World, standing at 42 per cent and Division II, The Mind, at 14 per cent. Emeritus Professor M.L. Samuels, the founder of the project, is working his way steadily through the enormous task of proof-reading completed sections.

Major completions included the Animal Kingdom, the largest category yet at 42,571 records. Work is in progress on two other major categories, Movement and Plants, on other sections remaining in the Material Universe section, and on the Probability section of Mental Activity. Pre-classification, ie the preparation of sections for the editors, continues to focus on Endeavour, one of the last remaining major sections.

Major grants from the Carnegie Foundation for the Universities of Scotland and the British Academy have eased financial problems considerably (but only for the time being). Smaller grants have enabled us to employ postgraduate students to help with editorial tasks. Work on the development of the Ingres database and an SQL front-end have been greatly facilitated by the purchase of a Sun work-station to act as an independent server.

The project was demonstrated at the Tenth International Conference on English Historical Linguistics at the University of Manchester in August 1998 as part of a Brook Symposium on 'The Oxford English Dictionary and English Historical Lexicography'; other projects demonstrated included the new OED,

the Dictionary of South African English, the Middle English Compendium, and the Thesaurus of Old English. A volume of papers is in preparation.

An Institute for the Historical Study of Language has been set up in the English Language Department at Glasgow to provide a forum for projects such as the Historical Thesaurus and the more recent Middle English Grammar project, run by Dr Jeremy Smith.

Christian Kay:

c.kay@englang.arts.gla.ac.uk

The Early Modern English Dictionaries Database (EMEDD) project: Update

The EMEDD is a foundation for a future dictionary of Early Modern English (ca 1450 – ca 1700). It consists of a full-text database of early bilingual (English-French, English-Italian, English-Latin, and English-Spanish) and monolingual English dictionaries, which hold the direct testimony of native English lexicographers. Some 200 EME lexicographic works, including enhanced re-editions, revisions of these, and reprints, exist up to the early 18th century. About 128,000 word-entries from eleven representative lexicons by John Palsgrave (1530), William Thomas (1550), Richard Mulcaster (1582), Edmund Coote (1596), John Minsheu (1599), Robert Cawdrey (1604), John Bullokar (1616), Henry Cockeram (1623), Thomas Blount (1656), and John Garfield (1657) are available publicly at the Renaissance Electronic Texts Web site <www.library.utoronto.ca/www/utel/ret/ret.html>. A Pat-based search engine enables researchers to obtain between 25 and 100 citations for any word, partial word, or word combination in all or in any one of these eleven dictionaries.

Work nears an end on proofing and processing bilingual dictionaries by Sir Thomas Elyot (Latin-English, 1538), John Florio (Italian-English, 1598), and Randle Cotgrave (English-French, 1611), and monolingual glossaries by John Aspley (1605), Robert Barrett (1598), Robert Cawdrey (1617), and Richard Rowlands (1605). When added to the current EMEDD by the year 2000, the total number of word-entries will almost double, to about 225,000.

Recently the following EMEDD-related research has been published. Renaissance Electronic Texts (RET) and Representative Poetry On-line (RPO) both brought out EMEDD-related editions. For 1997, RET included Edmund Coote's *The English Schoole-maister* (1596), and for 1998, RPO offered George Puttenham's *The Arte of English Poesie* (1589). A special issue of *Early Modern Literary Studies* in April 1997, edited by Michael Best and myself at <purl.oclc.org/emls/emlshome.html> consists of essays on *New Scholarship*

the form of both word and ‘dot’ maps. It is likely that a number of different types of map will have to be employed to display the data when the corpus is finally completed. Some areas of the country have very sparse coverage of available material or are quite simply empty of data. Others are too crowded for comparable levels of display to be feasible. It is likely that LAEME will include word maps, comparable with those in LALME (*Linguistic Atlas of Late Middle English*), for the whole country for only a few items. Dot maps, perhaps in some cases displaying more than one feature, as contemporary mapping software now allows, will be presented for many more items. Some areas – certainly the South-West Midlands – will require larger-scale presentation in order for the data to be clearly readable. Producing series of maps for specific areas will avoid the problem posed for the early Middle English material, that considerable areas of the map would otherwise be empty.

Associated research

The slight slowing down of the tagging process this year compared to previous years has coincided with more emphasis on mapping and data display and with more exploration of the power of the tagging method in revealing linguistic patterns and understanding better the linguistic behaviour of scribes. Much less early Middle English survives than late Middle English, and the texts that do survive are often not ideal witnesses. But we cannot afford, as we could with LALME, to discard texts in mixed or recalcitrant forms of language. Some of the largest and best known examples of early Middle English survive only in texts in mixed or recalcitrant forms of language. Some studies of individual manuscripts have already been published and others are in progress and forthcoming. These are important stages towards the making of LAEME. Linguistic complexity makes detailed scrutiny necessary in order for different language types to be correctly identified and sorted. Only then is it possible to say where the types of language belong. Other studies (see publications below) illustrate how the project methodology is a discovery procedure not only for data about regional linguistic variation and dialect mapping. Data from the tagged texts have led to observations about early Middle English writing systems, new textual readings, interpretations and datings, and hitherto unrecorded words, as well as providing information for the study of historical syntax.

Recent and forthcoming publications

1998. Linguistic and textual relationships between the Corpus, Nero and Vernon Manuscripts of Ancrene Riwe – a response. *Medieval English Studies Newsletter* 38: 4–16.
1998. Raising a stink in *The Owl and the Nightingale*: a new reading at line 115. *Notes and Queries* 243: 276–84.
1998. Notes on Oxford, Bodleian Library, MS Digby 86, the names of a hare in English. *Medium Aevum* 67: 201–11.
1998. Three notes on Dame Sirith, Oxford, Bodleian Library, MS Digby 86, fols. 165r–168r. *Neuphilologische Mitteilungen* 99: 401–409.
- 1999a (in press). ‘Never the twain shall meet’. Early Middle English the east west divide. In I. Taavitsainen, T. Nevalainen, P. Pahta and M. Rissanen (eds) *Language and Text: Selected Papers from the Second International Conference on Middle English* (Helsinki, 1997). Berlin: Mouton de Gruyter.
- 1999b (forthcoming). Confusion wrs confounded: litteral substitution sets in Early Middle English writing systems. *Neuphilologische Mitteilungen*.
- 1999c (forthcoming). The linguistic stratification of the Middle English texts in Oxford, Bodleian Library, MS Digby 86.
- In preparation: Negation in Early Middle English.

List of Tagged Texts in the Early Middle English Corpus that have been added since the last ICAME report

- Cambridge, Corpus Christi College 8, p 457: *Worldes blisce haue god day*.
- Cambridge, Gonville and Caius College 52/29, fol. 43r: *Creed, Pater Noster, Ave Maria*, In *Manus Tuas* Cambridge University Library Hh.6.11, fol. 70v: *Pater Noster and Ave Maria*.
- Dublin, Trinity College 432, fol. 22r: *My leman on the rood*.
- Hereford Cathedral Library, O.III.11, fol. 122v: *lyric on the Passion*.
- London, British Library, Additional 25031, fol. 5v: *Ten Commandments*.
- London, British Library, Cotton Faustina Av, fols. 10r–v, 105v–106r: *short verses*.

London British Library, Cotton Vitellius, D.iii, fols. 6r–8v: fragments of Floris and Blanche-flur.

London, British Library, Royal 2.F.viii, fol. 1v: two lyrics.

London, British Library, Royal 12.E.i, fols. 193r–194v: three lyrics in two hands.

London, British Library, Royal 17.A.xxvii, fols. 1r–70v: The Katherine Group in 3 different hands.

Oxford, Bodleian Library, Digby 86, fols. 168r–v, 195r–200r, 206r: 6 verse texts not previously tagged.

Oxford, Corpus Christi College 59, fols. 66r–v, 113v, 116v: verses on God and the BVM.

Salisbury Cathedral, MS 82, fol. 271v: Pater Noster.

Margaret Laing: m.laing@ed.ac.uk