English historical corpora: Report on developments in 1997

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) and 21 (1997), and in *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). Further information on corpus-based work done within the diachronic approach can be found in a recent volume titled *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 workshop organized in Chester prior to the 18th ICAME conference in May 1997. The next workshop on this topic will be arranged in Newcastle (Northern Ireland) 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

PROJECT COMPLETED

A Thesaurus of Old English, by Jane Roberts and Christian Kay with Lynne Grundy, was published by King's College London, late in 1995. As well as being a pilot project for the Historical Thesaurus, the Thesaurus of Old English is being acknowledged as a significant addition to resources for scholars of Old English. Copies of the two-volume

work can be obtained from Professor David Hook, Spanish Department, King's College London, Strand, London WC2R 2LS (price GBP 57.50).

(*Corpora Across the Centuries*, pp 11–120, 155–161;
ICAME Journal 19: 152–153; 20: 126; 21: 118)
Jane Roberts: J.Roberts@kcl.ac.uk

HISTORICAL CORPORA

The Brooklyn-Geneva-Amsterdam-Helsinki Parsed Corpus of Old English
The corpus project aims at a glossed, morphologically tagged, and syntactically tagged and bracketed version of the prose texts of the Old English section of the Helsinki Corpus. The annotation will eventually be extended to cover the entire Toronto Dictionary of Old English corpus.

Two groups of scholars from three countries are collaborating on the project. The first group includes Ans van Kemenade, Willem Koopman, and Frank Beths (Amsterdam, the Netherlands), and is responsible for the morphological tagging of the corpus; the second group includes Susan Pintzuk (York, England) and Eric Haeberli (Geneva, Switzerland), and is responsible for glossing, syntactic tagging and bracketing, and the information retrieval and data manipulation programs. 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. The programs to gloss and partially automate the syntactic tagging and bracketing have been completed and are being used to produce glossed and syntactically annotated text. The programs for information retrieval and data manipulation have been designed, and will be written and implemented before the end of 1998. The corpus is expected to be in distribution within two years.

(*ICAME Journal* 19: 151; 21: 112–113)
Susan Pintzuk: SP20@york.ac.uk

Corpus of Early English Medical Writing 1375–1750

The Corpus of Early English Medical Writing is currently being developed at the English Department of the University of Helsinki by Irma Taavitsainen and Päivi Pahta. The plan covers the period from 1375 to 1750. The medieval part of the corpus extends to 1550, and is further divided into two periods, 1375–1475 and 1475–1550. The selection of pre-1475 texts is nearly completed, but work on the parts after that date is still in progress. At present the medieval part of the corpus contains ca 535,000 words which includes surgical and anatomical texts, special fields, encyclopaedias and compendia, remedybooks and recipes, and medical verse. Shorter texts are included in all, and longer treatises are represented by extracts of at least 10,000 words. At present the corpus is not available for public use.

For the plan, see *ICAME Journal* 21, 1997, pp 71–78, and for the most recent introduction, see *Medieval English Studies Newsletter* (University of Tokyo), 1998 forthcoming.

Irma Taavitsainen: Irma.Taavitsainen@helsinki.fi
Päivi Pahta: Paivi.Pahta@helsinki.fi

The Lampeter Corpus of Early Modern English Tracts

The news concerning the Lampeter Corpus is available in the internet address http://www.tu-chemnitz.de/~ehe/real/real.htm

(*Corpora Across the Centuries*, pp 81–89;
ICAME Journal 19: 151–152; 21: 110)

Josef Schmied: Josef.Schmied@phil.tu-chemnitz.de
Claudia Claridge: Claudia.Claridge@phil.tu-chemnitz.de
Rainer Siemund: Rainer.Siemund@phil.tu-chemnitz.de

Edinburgh Corpus of Older Scots
Institute for Historical Dialectology, University of Edinburgh

Keith Williamson is compiling a lexicogrammatically tagged corpus of Older Scots texts (ca 1380 to 1500). The purpose of the corpus is to provide a data-base for the study of linguistic variation in Medieval

Scots texts. The corpus will be used first to examine diatopic and diachronic variation in Early Scots and Early Middle Scots, with the first part of an historical linguistic atlas of Older Scots as principal objective. This will include investigation of the linguistic relationships between coeval Scots and Northern English texts. The Edinburgh Corpus of Older Scots is a partner project of the Linguistic Atlas of Early Middle English, both deploying the same methodology.

Keith Williamson: I.K.Williamson@ed.ac.uk

Leeds Corpus of English Dialects

This corpus project, carried out by Juhani Klemola at the University Leeds, aims at an orthographically transcribed machine-readable corpus of traditional dialect speech. The corpus consists of 314 recordings from 289 localities in the SED network, totalling approximately 60 hours of speech. The size of the corpus is about 700,000 words.

At the time of writing, we have completed the first stage of the project, the first version of the orthographic transcriptions. We are currently working on checking and proof-reading the transcripts. The project is funded by a Leverhulme Trust grant (F/122/AT; January 1997-August 1998).

(*ICAME Journal* 21: 111–112)
Juhani Klemola: J.Klemola@leeds.ac.uk

HISTORICAL DICTIONARIES AND ATLASES

Dictionary of Old English Project

The Dictionary of Old English project is very pleased to announce the launch of its latest research tool. The Dictionary of Old English Corpus in Electronic Form, used in the preparation of the *Dictionary of Old English*, is now accessible for searching on the World Wide Web. This online database includes at least one copy of every Old English text

(and sometimes more than one copy, if significant because of dialect or date). It represents over 3,000 texts and contains some three million running words of Old English and another two million running words of Latin. The texts of the searchable Corpus are SGML-encoded and fully conformant with the 1994 Guidelines for Electronic Text Encoding and Interchange (TEI P3). The Dictionary of Old English Corpus, therefore, is one of the few corpora in the world to conform fully with what has become the standard for electronic text. Accessing the database (1997 version) via an interface developed by John Price-Wilkin of the Humanities Text Initiative (HTI) at the University of Michigan, users will be able to search both Old English and Latin for spellings, single words and phrases, and do boolean and proximity searches, in each case viewing the results in a three-sentence context.

As an aid to designing useful searches, we have developed with HTI a full index of every spelling in the Corpus divided into OE and Latin word wheels. These are further subdivided into letters of the alphabet within each language. Thus, a user can choose to scroll through, for example, all the spellings in Old English *F* in order to check for all possible spellings of the adjective *fæger*. Once these are bolded, they constitute a 'hit' list, and the user can then click on each of these spellings in turn to bring up their contexts in the Corpus. We are delighted at having so easily accessible an index of spellings in the Corpus which can then be expanded into full-sentence contexts. The word wheels add both efficiency and comprehensiveness to the design of our searches because they provide online the range of possible spellings.

The Web Corpus is available to institutions by subscription for a yearly access fee of $200.00. At this time, the Corpus is available only to institutions with an internet IP address; that is, individuals gain access only through formal affiliation with a subscribing institution. In the normal case, authorized users at a site must be employees, faculty, staff, or students at the institution. The Web Corpus is a publication of SEENET (Society for Early English and Norse Electronic Texts), a series under the general editorship of Hoyt Duggan, University of Virginia. Distribution is handled by the University of Michigan Press. The subscription agreement and information on the Web Corpus can be found at the following address:

http://www.press.umich.edu/webhome/healey/siteform.html

We are pleased at this first step in our plan to build a distributed

archive of our research material in early English.

(*ICAME Journal* 21: 116–117)
Antonette diPaolo Healey: Healey@doe.utoronto.ca
For inquiries about the
Electronic Corpus: Corpus@doe.utoronto.ca

Linguistic Atlas of Early Middle English
Institute for Historical Dialectology, School of Scottish Studies, University of Edinburgh

The Corpus of Early Middle English Tagged Texts and Maps
The main task of transcribing and tagging more early Middle English texts has continued. The corpus of early Middle English texts transcribed and tagged now consists of 249 texts from 78 different manuscripts of which 48 (from 17 different manuscripts) have been added since the last report. (See the list below.) These texts vary in length but average 1,047 words. To date 327,995 words of text have been tagged, 50,277 since the last report. From the tagged corpus dictionaries are generated which to date contain 27,947 different tags describing 44,951 different forms. The tagged corpus now represents 113 different hands or types of early ME of which 75 have been given provisional placings.

Associated Research
We have given a great deal of thought and attention this year to theoretical aspects of the early Middle English Atlas project and the way forward towards its eventual completion. As the work has progressed, the methods of analysis we are using have been shown to be of considerable power and value in many unexpected ways. We are more convinced than ever that the methodology is sound and that the display of the early Middle English data in the form of an Atlas is not only possible, but also the best and most sensible way of presenting its complexity. What is even more exciting, however, is how much we are discovering about the detail and variability of written English at this particularly fast-changing period and about the scribal policies which form the surviving textual sources. Data from the tagged texts has been used to make observations about early Middle English writing systems,

new texual readings, and hitherto unrecorded words. This sort of work takes time from the main task of tagging and mapping texts, but has to be done in order for different language types to be correctly identified and sorted. It is also of considerable interest in the wider field of early ME studies.

As more publications come out from the work of the project, there has been increasing interest from the wider community of ME scholarship. The many enquiries have included questions about (1) localisation of texts; (2) historical lexicography; (3) identification of 'new' ME texts; (4) diachronic text corpora and computers; (5) history of language and scribal copying practices; (6) language variation, textual readings and manuscript stemmata.

Recent and Forthcoming Publications

1997. A Fourteenth-Century Sermon on the Number Seven in Merton College, Oxford, MS 248, *Neuphilologische Mitteilungen* 98: 99–134. Helsinki.

1998 (forthcoming). 'Never the twain shall meet'. Early Middle English – the east west divide. In Irma Taavitsainen et al eds, 'Proceedings of Language and Text, the Second International Conference on Middle English', Helsinki, 1997.

1998 (forthcoming). Raising a Stink in The Owl and the Nightingale: a New Reading at line 115. *Notes and Queries.*

1998 (forthcoming). Confusion wrs Confounded: Litteral Substitution Sets in Early Middle English Writing Systems.

List of Tagged Texts in the Early Middle English Corpus that have been added since the last ICAME report (1997)

Cambridge, St John's College 111 (E8), fol. 106v: Stand wel moder

Cambridge, Trinity College B.14.39 (323), hand A: fols. 19r, 25v, 27rb, 28r–29v, 32r–33v, 36r–46r, 47r–v, 83v–84r; hand B: fols. 20r–25r, 26r–27rb, 27v, 34r, 35r–v; hand C: 30r–31v, 81v; hand D: fols. 81v–82r, 85r–87v. Verse texts including Proverbs of Alfred.

Cambridge University Library Add 3020, Red Book of Thorney 1, fol. 18r: Will of Mantat

Cambridge University Library Add 3021, Red Book of Thorney 2, fol. 372: Document, Kingsdelf

London, British Library, Additional 27909, fol. 2r: Penitence for wasted

life
London, British Library, Cotton Caligula A ix, fols. 246r–249r: seven verse texts
London, British Library, Cotton Charter iv.18: Athelstan's Charter to Beverley
London, British Library, Cotton Cleopatra B vi, fol. 204v: Pater Noster, etc.
London, British Library, Cotton Julius A v, fols. 180r–181v: Poem on the Scottish Wars
London, British Library, Cotton Nero A xiv, fols. 1r–11r: Ancrene Riwle, Part 1
London, British Library, Cotton Vitellius A xiii, Chertsey Cartulary, fols. 50r–51v, 53v: six documents
London, British Library, Egerton 613, hand A, fol. 1v: Somer is comen; hand B, fol. 2r: Stella Maris; hand C, fol. 2r–v: Blessed be thu lauedi; hand D, fol. 2v: Litel uoit eniman
Oxford, Bodleian Library, Digby 86, fols. 119r–143r, 163v–168r: 14 verse texts (six still to do)
Oxford, Bodleian Library, Rawlinson G.18, fols. 105v–106v: lyric on the vanity of the world
Oxford, Jesus College 29, fols. 179v–185v: six verse texts
Oxford, New College 88, fols. 31r, 179r–v, 488v: lyrics and 10 Commandments
Worcester, Herefordshire and Worcestershire Record Office, BA 3814, fol. 38v: Document, Worcester

(*Corpora Across the Centuries*, pp 121–141;
ICAME Journal 19: 154–155; 20: 126–130; 21: 119–120)
Margaret Laing: M.Laing@ed.ac.uk