English historical corpora: Report on developments in 1996

Merja Kytö and Matti Rissanen
Uppsala University and University of Helsinki

After the First International Colloquium on English Diachronic Corpora, held in March 1993 at St Catharine's College, Cambridge, Merja Kytö and Matti Rissanen have chaired historical corpus workshop sessions arranged on the occasion of recent ICAME Conferences (Zürich 1993, Aarhus 1994, Toronto 1995). In May 1996, preceding the 17th ICAME Conference, a two-day workshop took place in Helsinki and Stockholm, and on board the ferry between the two cities. The next workshop will take place on 20–21 May, immediately prior to the 18th ICAME Conference due in Chester.

Reports on the year's work on English historical corpora, thesauruses, atlases and dictionaries have been published in *ICAME Journal* (1995, 19: 145–158; 1996, 20: 117–132). The proceedings of the Toronto workshop will appear in *Tracing the Trail of Time*, edited by Raymond Hickey, Merja Kytö, Ian Lancashire and Matti Rissanen (Rodopi, 1997).

The present report will supplement those included in *Corpora Across the Centuries: Proceedings of the First International Colloquium on English Diachronic Corpora* (edited by Merja Kytö, Matti Rissanen and Susan Wright, Amsterdam and Atlanta, GA: Rodopi, 1994) and the reports published in *ICAME Journal* 19 and 20. Each entry below is followed by references to those reports.

We thank 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

A CORPUS COMPLETED

1 The Lampeter Corpus of Early Modern English Tracts

The Lampeter Corpus – its name derives from the Founders' Library at the University of Wales Lampeter from which the corpus material was taken – comprises short tracts published between 1640 and 1740. The collection is made up of 12 texts per decade, altogether 120 different texts by 120 different authors, which are subcategorised into the domains 'economy/trade', 'politics', 'religion', 'law', 'science' and 'miscellaneous'.

Both the contents and style of the texts mirror the range of contemporary non-fictional prose publications. To allow for textlinguistic research, only complete texts were included, altogether amounting to some 1.1 million words.

The textual markup of the corpus further provides information on the structural and layout characteristics of the texts by using TEI/SGML conformant coding. Text headers accommodate chiefly extralinguistic pointers to the background of authors, printers/publishers, print places or text types. Altogether, the corpus serves as a powerful tool to explore the highly influential production type of the short tract or pamphlet at a time that marks the rise of both mass production of printed matter and mass literacy. It is currently available through ICAME and the Oxford Text Archive and will be published in a part-of-speech-tagged version at a later stage. For a more detailed description of the corpus, see this volume of the *ICAME Journal* or contact the compilers at the University of Technology at Chemnitz, Germany. The Lampeter project is funded by the *Deutsche Forschungsgemeinschaft*, the German Research Foundation.

(*Corpora Across the Centuries*, pp 81–89;
ICAME Journal 19: 151–152)
Josef Schmied: Josef.Schmied@phil.tu-chemnitz.de
Claudia Claridge: Claudia.Claridge@phil.tu-chemnitz.de
Rainer Siemund: Rainer.Siemund@phil.tu-chemnitz.de

NEW CORPUS PROJECTS

2 Corpus of Early English Medical Writing 1375–1750

This new project focuses on the evolution of medical writing within the variationist framework of stylistics and discourse analysis. The corpus under compilation will serve as material for the compilers' research project *Scientific thought-styles: The evolution of English medical writing.* When completed, the corpus will consist of c. a million words; the current version contains c. 300,000 words. In the first phase the material is drawn mainly from the Late Middle English and Early Modern periods. For an introduction to the project, see the article by Taavitsainen and Pahta in this issue of the *ICAME Journal.*

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

3 The Corpus of Women's Scots

Anneli Meurman-Solin (University of Helsinki) is compiling a computer-readable corpus of Scottish women's early writings. The texts date from 1540–1800 and represent genres such as private and official letters, autobiographical writings, essays on various topics, travelogues and drama.

Anneli Meurman-Solin: Anneli.Meurman-Solin@helsinki.fi

4 Leeds Corpus of English Dialects

Juhani Klemola (University of Leeds) is currently working on a project aiming at a corpus of traditional dialect speech. The material consists of tape-recordings made in the 1950s and early 1960s in connection with the Survey of English Dialects project. The surviving tape-recordings

from c. 250 SED localities are relatively short, about 8 minutes on average, but the total length of the recordings still adds up to c. 35 hours of traditional dialect speech. We estimate that the transcribed corpus will consist of about 700,000 words.

Our objective is to produce a corpus that consists of orthographically transcribed text and sound files of the actual tape- recordings aligned. The work on the first stage of the project, the orthographic transcription of the recordings (carried out by research assistant Mark Jones), started in January 1997. The project is funded by a Leverhulme Trust grant (January 1997 – August 1998).

Juhani Klemola: J.Klemola@leeds.ac.uk

PROGRESS OF EARLIER PROJECTS

5 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 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 is 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 1997. The corpus is expected to be in distribution within three years.

(*ICAME Journal* 19: 151)
Susan Pintzuk: SP20@york.ac.uk

6 Penn-Helsinki Parsed Corpus of Middle English

This corpus project, carried out by Anthony Kroch and Ann Taylor (University of Pennsylvania), contains over half a million words of syntactically annotated Middle English made up from the Middle English prose section of the Helsinki Corpus plus some additional texts. The annotation consists of labelled brackets which indicate a combination of function and form making automatic searching of syntactic constructions possible. The documentation and utilities files for the corpus are freely accessible via:

anonymous ftp
babel.ling.upenn.edu/research-material/mideng-corpus
gopher University of Pennsylvania Linguistics Department
babel.ling.upenn.edu (port 70)
World-Wide Web
http://www.ling.upenn.edu/mideng/

The texts themselves are available to registered users. Details on how to register are contained in the README file at the above-mentioned site.

Phase II of this project, now underway, involves (1) part-of-speech tagging of the existing corpus, (2) tagging and parsing the poetry section of the Helsinki Corpus, and (3) enlarging the prose section of the corpus by entering, tagging and parsing at least another half million words of text.

(*ICAME Journal* 19: 157)
Anthony Kroch: kroch@change.ling.upenn.edu
Ann Taylor: ataylor@linc.cis.upenn.edu

7 The Corpus of Early English Correspondence (CEEC)

The Corpus of Early English Correspondence (CEEC) is compiled for historical sociolinguistic research and constitutes the main data source for the *Sociolinguistics and language history* project currently under way at Helsinki University. The 1996 version of the corpus covers the period 1420–1681 and consists of c. 2.5 million running words. The first version of the CEEC based on personal files was also completed in 1996. This format consists of separate files for each individual writer with more than 2,000 running words. The majority of the writers included in the corpus are represented by more than 2,000 words and can now be searched individually.

The corpus team has lately concentrated on the social representativeness of the corpus checking all socially underrepresented subperiods in the CEEC. A large number of Early Modern English letter collections were consulted in 1996 and new material from thirty editions was selected for inclusion in the corpus. Processing this additional material will continue in 1997, as will the second proofreading of the corpus, which began in the autumn of 1996.

As part of the proofreading process, the corpus team has checked the oldest editions included in the corpus against their manuscript originals. This work, carried out in the British Library, the Public Record Office and Dulwich College Library, yielded highly satisfactory results in that most of the tens of collections checked proved to be quite reliable and certainly up to the standard required in morphological and syntactic studies. Those few that were found less satisfactory were carefully checked, and new, corrected versions of them will be included in the 1997 version of the CEEC.

In 1996 the team published their first joint publication (T. Nevalainen and H. Raumolin-Brunberg (eds), *Sociolinguistics and Language History; Studies based on the Corpus of Early English Correspondence*. Amsterdam: Rodopi). The articles in the volume first introduce historical sociolinguistics and the research material used, the CEEC, and then proceed to testing the role in language change of such social variables as gender, age and social status (Nevalainen, Raumolin-Brunberg). Early standardization of English is also discussed (Kirsi Heikkonen), as are certain individual processes of change, including the epistemic parenthetical METHINKS (Minna Palander-Collin), periphrastic DO and BE + ING (Arja Nurmi) and forms of address (Helena Raumolin- Brunberg).

(*ICAME Journal* 19: 147; 20: 124)
Terttu Nevalainen: Terttu.Nevalainen@helsinki.fi
Helena Raumolin-Brunberg: Helena.Raumolin-Brunberg@helsinki.fi

8 A Corpus of Dialogues, 1550–1750

Over the last year, Jonathan Culpeper (Lancaster University) and Merja Kytö (Uppsala University) have been collaborating on a project aiming at a corpus of texts reflecting spoken dialogue from 1550 to 1750. Whilst our overall plan is to construct a corpus of a good million words, as a first step we decided to build a pilot corpus of 360,000 words, divided equally between four text types – trial proceedings, witness depositions, drama, and prose fiction – taken from the period 1590 to 1720. Thus, half of the pilot corpus would contain naturally occurring speech (supposedly recorded verbatim or nearly so) and half constructed imaginary speech. Furthermore, half (trial proceedings and drama) would be recorded with minimal explicit narratorial interference and half (witness depositions and prose fiction) with considerable interference.

With the help of a grant from the British Academy, work began on the pilot corpus in June 1996. At the time of writing, approximately 250,000 words are in electronic form. As might be imagined, we have found some difficulty in finding suitable texts. Locating suitable drama texts was relatively easy, in spite of the fact that much drama is written in verse – something we had determined to avoid. Records of trial proceedings were also relatively easy to find, though they are not found in great abundance prior to the 18th century. Finding suitable witness depositions and prose fiction extracts is proving troublesome. The main problem here is finding texts which contain sufficient quantities of reported speech. With some exceptions, collections of witness depositions have not proved easy to find, and some depositions are more summaries of a speech event rather than close reports of it: they owe much more to the recording clerk than the original speaker. Much well known prose fiction contains no speech presentation at all. We hope to find suitable extracts in minor works.

In the process of constructing the pilot corpus, we have become increasingly aware that two areas are in need of further development. Firstly, it is clear that in order to attempt to interpret the dialogue considerable amounts of contextual information are needed (about the

participants, the speech event and so on). This may argue for the future construction of some kind of database. Secondly, it is clear that we need to develop a more sophisticated and systematic coding system to enhance the usability of the corpus. This is necessary if we wish, for example, to compare the speech of male speakers with that of female, or if we are to attempt to identify the speaker's contribution to the text, as opposed to the explicit narratorial contribution.

(*ICAME Journal* 20: 121–122)
Jonathan Culpeper: J.Culpeper@lancaster.ac.uk
Merja Kytö: Merja.Kyto@engelska.uu.se

HISTORICAL DICTIONARIES AND ATLASES

9 Dictionary of Old English Project

The *Dictionary of Old English Corpus in Electronic Form* contains at least one copy of every surviving Old English text, including poetry, prose, glosses, glossaries, runic, and non-runic inscriptions. It consists of three million running words of Old English and two million running words of Latin, 3,025 texts in all, and occupies with overhead 38MB. The most recent version of the Corpus (1995, on diskette) is among the first humanities databases to be fully conformant with the 1994 Guidelines issued by the *Text Encoding Initiative (TEI) P3*, edited by Lou Burnard and Michael Sperberg-McQueen. The TEI-P3-conformant coding was implemented through the collaborative efforts of Takamichi Ariga of the Dictionary staff and John Price-Wilkin of the University of Michigan.

A Web interface was developed for the Dictionary's Electronic Corpus by John Price-Wilkin in 1995. It uses PAT as its search engine, the powerful tool developed by the University of Waterloo (Canada) for the Electronic Version of the *OED*, and now distributed by Open Text Corporation. The Web Corpus is available presently only to the University of Toronto community. However, negotiations are underway with a major university press to make the Web Corpus, which displays the results of

sophisticated searches, generally available either by site license or by individual subscription.

The updating and correcting of the Electronic Corpus is an ongoing task. We are now updating the editions contained in the Corpus as recent work in the field is published. A list of the new editions which are currently being input can be found in the *Preface* to *Dictionary of Old English: E* (1996). Also as a result of the citation check which is associated with the publication of each fascicle of the *Dictionary*, corrections are entered not only in the fascicle, but also in the source of its quotations, the Electronic Corpus. The maintenance of these electronic databases is crucial so that they do not become fossils.

In preparation for the publication of the next letter of the *Dictionary*, *F*, which we hope to publish on CD-ROM together with the previous six (*A, Æ, B, C, D,* and *E*), we are trying to normalize fully the way in which we refer to the Latin sources to the Old English texts. We have adopted the system devised by Michael Lapidge, University of Cambridge, for the Fontes Anglo-Saxonici and the Sources of Anglo-Saxon Literary Culture Projects, and later supplemented by Pauline Thompson of the Dictionary staff, for many anonymous sources. As Lapidge's *Abbreviations of Sources* (1988) was published after our earliest fascicles, and is now undergoing further revision, we are eager to incorporate the latest changes throughout the *Dictionary*. By their inclusion we hope to encourage a standard system of reference for Latin sources to Old English texts. Papers, written by Antonette diPaolo Healey and Nancy Speirs, which discuss more fully the corpora of the Dictionary of Old English Project will be published in the proceedings of the ICAME 1995 volume.

Antonette diPaolo Healey: Healey@doe.utoronto.ca
For inquiries about the Electronic Corpus: Corpus@doe.utoronto.ca

10 Progress on the Historical Thesaurus of English 1996

Contrary to rumours circulating in some places, the Historical Thesaurus of English project has not yet been published. (The one that has been published is A Thesaurus of Old English, Jane Roberts and Christian Kay with Lynne Grundy, King's College London Medieval Studies XI, 1995). However, we continue to edge towards that desirable state, with increasing amounts of data being classified and entered in the database. Sections added during 1996 include Possession (11,017 records, with Take by far the largest subsection at 3,615); Authority, including sub-sections such as Politics and Punishment (25,854 records) and Animals (currently standing at 30,000 records but not yet complete). Work is in progress on Mind, Ships, Maths and Chemistry.

This leaves various scientific sections still to be classified, plus the two major sections on Endeavour and Existence.

The calendar year 1996 was in fact a record one for data entry, bringing the total number of records held to around 470,000. Professor Emeritus Michael Samuels, the founder of the project, has begun the mighty task of proofreading the work, and online corrections are being made. Further refinements have been made to the Ingres database, and work is progressing on a user-friendly front end to cover the most common queries.

Funding has become a major preoccupation again, with our Leverhulme Trust Grant coming to an end in September. The British Academy has awarded us a Larger Research Grant, which will support a key salary in 1997–1998, but we have some way to go in finding other essential support.

(*Corpora Across the Centuries*, pp 111–120, 155–161;
ICAME Journal 19: 152–153; 20: 126)
Christian J. Kay: CJKAY@human.gla.ac.uk

11 Middle English Word Studies

This project, co-authored by Jane Roberts and Louise Sylvester, will offer new resources to scholars in the fields of lexicology, history of the language, and the literature and cultural history of the period, as

well as providing essential work towards a 'Middle English Thesaurus'. The proposed first volume, 'Middle English Word Studies; A Word and Author Index', will contain detailed bibliography specifically about Middle English vocabulary and will be a research tool for new work on the lexis of the period, for example: lexical field studies; loan words; vocabulary loss; dialect diversity. The second volume, 'Middle English Semantic Field Studies', will examine the way in which the vocabulary of the Middle English period grew and structured itself and show which lexical fields have been examined and which are as yet unexplored. For further information, see Louise Sylvester and Jane Roberts, 'Middle English Word Studies', *Medieval English Studies Newsletter* 34 (1966), 8–11.

Jane Roberts: J.Roberts@kcl.ac.uk

12 Linguistic Atlas of Early Middle English, Institute for Historical Dialectology, University of Edinburgh

The Corpus of Early Middle English Tagged Texts and Maps

This year's effort has been devoted to the main task of transcribing and tagging more early Middle English texts and placing and mapping their processed language forms on provisional working maps. The corpus of early ME texts transcribed and fully tagged (for both meaning and grammatical function) now consists of 201 texts from 66 different manuscripts of which 38 (from 14 different manuscripts) have been added since the last report. (See the list below). The whole corpus is continually subject to correction and revision as the addition of further texts makes this necessary. To date 277,718 words of text have been tagged, 57,959 since the last report. From the tagged corpus dictionaries are generated which to date contain 25,023 different tags describing 39,120 different forms. The tagged corpus now represents 86 different hands or types of early ME language of which 65 have been given provisional placings on the map. New, more sophisticated mapping software makes it quicker to produce working maps, and the greater flexibility of presentation ensures that the material on them is now more

easily readable and comparable with the already published maps of the later Middle English language forms in the Linguistic Atlas of Later Mediaeval English. So far, 30 different 'item maps' have been produced as research tools to help in the placing of further texts. The next stage will be to identify suitable items for presentation as 'feature maps', comparable with the 'dot maps' in LALME.

List of Tagged Texts in the Early Middle English Corpus that have been added since the last ICAME report (1996)

Aberdeen University Library 154, fol. 368v: couplet and three quatrains
London, British Library, Cotton Nero A xiv, fols. 120v–131v: On God Ureison of ure Lefdi, Ureison of God Almihti, Lofsong of ure Lefdi, Lofsong of ure Louerde, Lesse Crede
London, Dulwich College XXII, fols. 81v–85v: La Estorie del Euangelie
London, Lambeth Palace Library 487, fols. 1r–59v, hand A: Lambeth Homilies; fols. 65v–67r, hand B: On Ureison of Ure Loverde
London, Westminster Abbey Library MS 34/3, fol. 36v: poem of impossibilities
Oxford, Bodleian Library, Ashmole 360, fol. 145v, hand B: lyric
Oxford, Bodleian Library, Ashmole 1280, fols. 48r, 192v: prayers
Oxford, Bodleian Library, Bodley 57, fol. 102v: lyric
Oxford, Bodleian Library, Digby 2, fols. 6r–v, 15r, 111r: lyrics
Oxford, Bodleian Library, Junius 121, fol. vi (flyleaf): Nicene Creed
Oxford, Merton College 248, ofls. 166r–157r: tags, lyrics and a sermon
Private: Blickling Hall, Norfolk MS 6864, fol. 35r: Creed
Worcester Cathedral, Chapter Library F.174, fols. 1r–66v: Ælfric's Grammar and Glossary, the Worcester Fragments.
Worcester Cathedral, Chapter Library Q.29, fols. 130v–131r: sermon

(*Corpora Across the Centuries*, pp 121–141;
ICAME Journal 19: 154–155; 20: 126–130)
Margaret Laing: Esss09@holyrood.ed.ac.uk