English historical corpora: Report on developments in 1999

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), 22 (1998) and 23 (1999). English historical corpora and the work done on them were the topic of the recent workshop organized in Freiburg (Germany) on the occasion of the 20th ICAME conference in May 1999.

We are grateful 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

PROJECTS COMPLETED

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 cor-

pus 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, is now available from Susan Pintzuk.

Susan Pintzuk: sp20@york.ac.uk
Ans van Kemenade: a.v.kemenade@let.kun.nl

The Penn-Helsinki Parsed Corpus of Middle English, Phase 2

Work is now complete on the The Penn-Helsinki Parsed Corpus of Middle English, Phase 2. This corpus project, carried out by Anthony Kroch and Ann Taylor at the University of Pennsylvania, contains over a million words of syntactically annotated Middle English made up primarily from extended samples of the texts included in the Middle English prose section of the Helsinki Corpus. The annotation consists of labelled brackets which indicate a combination of function and form making automatic searching for syntactic constructions possible. The new corpus contains a richer annotation than the first release, the PPCME1, including part-of-speech tags and internal structure for sentence constituents. In addition, the new corpus will be distributed with a linguist-friendly search engine, CorpusSearch, written specifically for the PPCME2 by Beth Randall. Official release of the PPCME2 is planned for mid 2000. Further information can be found at http://www.ling.upenn.edu/mideng/.

Anthony Kroch: kroch@change.ling.upenn.edu
Ann Taylor: at9@york.ac.uk

ICAMET: The Innsbruck Computer Archive of Machine-Readable English Texts

ICAMET, the Innsbruck Computer Archive of Machine-Readable English Texts, has been finished. As a full-text data base it encourages corpus analysis which presupposes complete text rather than extracts as a basis. The two parts of ICAMET, ie the Innsbruck Letter Corpus (1386 to 1688) and a sampler of the Innsbruck Prose Corpus (1100 to 1500), are now available as part of the ICAME CD-ROM (2nd ed), which can be ordered from the ICAME Archive of the University of Bergen. A CD-ROM of the Innsbruck corpora alone can at cost also be obtained from my department (Dept. of English, University of Innsbruck, Innrain 52, A–6020 Innsbruck). The complete version of the prose corpus can be used only from a CD-ROM in my department at Innsbruck, due to the well-

known copyright restrictions. I am still optimistic that these will soon be relaxed in cases of fair academic use.

The texts of the two parts of the ICAMET corpus have been encoded in DOS-WORD (only-text mode), with a few higher ASCII signs, which can easily be transferred manually in individual cases and automatically with the help of the keydriver 'ICAMET' added on the ICAME CD-ROM. Without the activation of this keydriver, some of the high ASCII signs are misrepresented on the ICAME CD-ROM, though always easy to decipher in their contexts.

A list of all corpus texts as well as the beginnings of both the prose books and the letters can be seen on my department's website (www.anglistik1.uibk.ac.at/ahp/staff/markus.html).

The details of the compiling policy are given in the Manual to ICAMET, which was published in 1999 and is available from my department or from any bookshop (ISBN 3–85124–163–0). Generally speaking, the corpus is close in format and layout to the editions used, except for diacritical signs; it is untagged and prepared for crunching in the sense that all editorial comments are marked to be excluded from the crunching process by programs such as WordCruncher Index. WC-Index lists of all corpus texts are planned for this year (2000).

While in their present sizes and make-ups the two corpora would no doubt profit from an expansion and from further editing work concerning, for example, parsing and tagging, no such work is intended. Given the present means and limits of my department and of myself, I invite colleagues of the international community to cooperate and perhaps to build up on what has been done.

Present and further work within the Innsbruck project on corpus linguistics refers to the semi-automatic normalisation of the ME and EModE texts, which are, though machine-readable, not fully accessible due to their extreme spelling variation.

Manfred Markus: manfred.markus@uibk.ac.at
http://www.anglistik1.uibk.ac.at/ahp/projects/icamet/index.html

The Early Modern English Dictionaries Database (EMEDD): Final update

In late 1999, the EMEDD was expanded to include 16 works. Its search engine, devised by Mark Catt, now draws on over 200,000 word-entries from 1530 to 1657. The collection includes bilingual and monolingual dictionaries and specialized glossaries and word-lists. There are six bilingual dictionaries: John Palsgrave's English-French (1530), William Thomas's Italian-English (1550), Thomas Thomas's Latin-English (1587), John Florio's Italian-English (1598),

John Minsheu's Spanish-English (1599), and Randle Cotgrave's French-English (1611). These cover four other languages and will yield pairs of French and Italian terms separated by 50–80 years. The six English-only (hard-word) dictionaries include well-known works by Edmund Coote (1596), Robert Cawdrey (1604 and 1617), John Bullokar (1616), Henry Cockeram (1623), and Thomas Blount (1656). Four specialized lexicons give more lexical variety to this mix: Bartholomew Traheron's translation of Vigon's work on medical terms (1543), William Turner's table of herbal names (1548), Richard Mulcaster's English word-list in his *The first part of the Elementarie* (1582), and John Garfield's scientific terminology in J. Renou's *Dispensatory* (1657).

The EMEDD can be accessed freely from the Renaissance Electronic Texts Web site at http://www.library.utoronto.ca/utel/ret/ret.html .

Recently agreement in principle was reached between the University of Toronto Press and the University of Toronto Library to publish electronically a Lexicon of Early Modern English (LEME), based on our past four years' experience with the EMEDD, but incorporating more research, an exhaustive library of texts, and much improved software. Research on this long-term, collaborative project will first turn to extend coverage to English printed and manuscript glossaries, dictionaries, and grammatical works for the period 1564–1616.

Scholars who are interested in this project are encouraged to contact me by e-mail at ian@chass.utoronto.ca or by post at New College, University of Toronto, Toronto, Ontario, Canada M5S 2Z3.

Ian Lancashire: ian@chass.utoronto.ca

NEW CORPUS PROJECT

The York-Helsinki Parsed Corpus of Old English

Anthony Warner and Susan Pintzuk of the University of York have received a three-year grant from the English Arts and Humanities Research Board to produce a parsed corpus of Old English. The principal researcher on the project, which began in January 2000, is Dr. Ann Taylor. The aim is to produce a 1.5 million word Old English corpus as a companion to the now complete Penn-Helsinki Parsed Corpus of Middle English, Phase II. The Old English corpus will be annotated using the same schema as the PPCME2 with some allowance made for the differences between Old and Middle English. The format of the two corpora will also be identical, making it possible to do automatic searching

on the Old English corpus using the linguist-friendly search engine, CorpusSearch, written specifically for the PPCME2 by Beth Randall.

Anthony Warner: aw2@york.ac.uk
Susan Pintzuk: sp20@york.ac.uk
Ann Taylor: at9@york.ac.uk

PROGRESS OF EARLIER 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
Leendert Plug: lp7@york.ac.uk

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

The objective of ECOS is twofold: (1) to compile a lexico–grammatically tagged corpus of diplomatically edited Older Scots texts; (2) to investigate linguistic variation in the texts. The first project is a study of diatopic and diachronic variation in texts written between c 1380 and 1500. Its aim is to make a linguistic atlas. The methodology employed is a development of that used for *A Linguistic Atlas of Late Mediaeval English*). This will be described in detail in Williamson (in preparation).

The work on extending the geographical matrix into North-east and South-west Scotland has been proceeding. This has entailed locating suitable manuscript sources as well as the transcription, keying and tagging of texts.

Some provisional maps of a few features were presented in a paper delivered at the 9th International Conference on Medieval and Renaissance Scottish Language and Literature (St Andrews, Scotland, 1999).

Wider issues of linguistic variation in Older Scots will be considered in collaborative studies with Anneli Meurman-Solin (Department of English, Univer-

sity of Helsinki). These will draw data from texts in the Helsinki Corpus of Older Scots and the Corpus of Scottish Correspondence as well as in ECOS. (See *ICAME Journal* no 23, p 179.)

A paper delivered at the Symposium on Linguistic Categories held at the University of Glasgow (15–17 September 1999) with Margaret Laing presented a precise method for unravelling and localizing strands in dialectal *Mischsprachen*. The method is based on the automated fitting algorithm developed for use in the Edinburgh historical atlas projects.

Other project-associated work in preparation

Prolegomena to a Linguistic Atlas of Older Scots [Report of research funded by The Leverhulme Trust, 1994–1998].

Borderline English – a linguistic comparison of 'Northern English' and 'Scots' of the late 14th and 15th centuries [For 11th International Conference of English Historical Linguistics, Santiago de Compostela, 7–11 September 2000].

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

Corpus of Early English Medical Writing 1375–1750

Work on the Corpus of Early English Medical Writing 1375–1750 was started in the English Department of the University of Helsinki by Irma Taavitsainen and Päivi Pahta a few years ago. The aim was to compile a computerised database for the project 'Scientific Thought-styles: The Evolution of Early English Medical Writing'. The project is funded by the Finnish Academy in 1999–2001, and the team has been joined by Martti Mäkinen, who works on his dissertation.

The medieval part of the corpus extends from 1375 to 1550, and it is further divided into two subperiods, 1375–1475 and 1475–1550. It contains c 535,000 words and is nearly completed now. Its text selection covers editions of medical texts from the first emergence of writings in this register in English, with aca-

demic tracts, surgical and anatomical treatises, texts in special fields like ophthalmology, encyclopaedias and compendia, remedybooks and recipes, and medical verse. Shorter texts are included in toto, and more comprehensive treatises are represented by extracts. Text selection for the Early Modern part of the corpus is under way, and preliminary work in charting the field has already been carried out. We aim to cover the widening use of English in medical writing in this period. At present the Early Modern part contains over half a million words, but it will grow in the future. The corpus is being tested with pilot studies by the project members, but it is not available for public use yet.

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

Corpus of Early English Correspondence (CEEC)

At the moment the *Corpus of Early English Correspondence* consists of approximately 2.7 million words covering the years 1410–1681. With the new funding in connection with the Research Unit for Variation and Change in English (VARIENG) at the Department of English, University of Helsinki, it has become possible to supplement the existing corpus in two ways. On the one hand, new editions of letters written in the time period covered can be included in the corpus. On the other hand, it also appears advisable to add to the time coverage of the CEEC: the project team are making plans for including letters from the eighteenth and nineteenth centuries. The principles for inclusion would follow mostly the same criteria as the existing CEEC, with emphasis on covering as wide a range of different social strata as possible. A particular emphasis will be on the language of women, since the increasing literacy of women makes this possible, and the role of gender has proved interesting in the development of most linguistic items studied. The new additions to the corpus are for the moment being called the CEEC Supplement, but it is possible that they will be fully incorporated into the CEEC. The sampler version of the corpus (CEECS) and its manual are available on the new ICAME CD-ROM; the manual can be found also on the ICAME web site.

More information at http://www.eng.helsinki.fi/doe/projects/ceec/ and *ICAME Journal* 23: 53–64. An updated list of collections can be found in the appendix of Palander-Collin (1999) *Grammaticalization and Social Embedding: I THINK and METHINKS in Middle and Early Modern English*. (Mémoires de la Société Néophilologique de Helsinki 55). Helsinki: Société Néophilologique,

and Nurmi (1999) *A Social History of Periphrastic DO*. (Mémoires de la Société Néophilologique de Helsinki 56). Helsinki: Société Néophilologique.

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
Minna Nevala: minna.nevala@helsinki.fi

A Corpus of English Dialogues 1560-1760

The aim of the project, currently under compilation at Lancaster University and Uppsala University, is to construct a computerised corpus of dialogue texts for research purposes. The general aim is to improve our understanding of the language of spoken face-to-face interaction in the Early Modern English period. Much of the research done on Early Modern English has been based on the language of literary and scholarly texts. Less emphasis has been placed on the use of dialogic speech-related texts. There is therefore an obvious need for a large computerised corpus focusing on speech-related data, offering easy access to a structured and contextualised selection of texts drawn from reliable sources.

The Corpus of English Dialogues will consist of both constructed dialogue – primarily drama comedies, handbooks or didactic works in dialogue form, and prose fiction – and records based on authentic dialogue, that is, trials and witness depositions. The text types drama comedies, handbooks and trials have minimum explicit narratorial interference, while prose fiction and witness depositions contain considerable explicit narratorial interference. The corpus will be divided into five forty-year subperiods, each containing a number of texts from each genre: the aim is to produce a computerised corpus of at least one million words; the corpus currently comprises ca 800,000 words.

The purpose of the corpus is to enable research into such fields as variationist studies, historical pragmatics, and stylistics. Relevant research questions are: What linguistic features characterised Early Modern English conversation? Did the dramatic texts of the time, including Shakespeare, accurately represent the characteristics of conversation? What role does the language of spoken face-to-face interaction play in language change? How have the pragmatic phenomena of conversation – with respect to both form and function – changed over the last 400 years? How was speech presented in writing?

The corpus compilers are Jonathan Culpeper (Lancaster University) and Merja Kytö (Uppsala University); other members of the team are Dawn Archer

at Lancaster University, and Mattias Jacobsson and Terry Walker at Uppsala University.

Jonathan Culpeper: j.culpeper@lancaster.ac.uk
Merja Kytö: merja.kyto@engelska.uu.se

ZEN nearing completion – ZENcomp under way

The first version of the Zurich English Newspaper Corpus, consisting of newspapers published between 1671 and 1791, is expected to be available in a CD-ROM format by spring 2001. Final revisions and extensions are currently being carried out. In its completed version the ZEN will include files ranging between 100,000 and 180,000 words for each of the decades investigated. This will mean that the total corpus may amount to more than 1,500,000 words.

To cover the entire period of early news publication in England, the ZEN team has decided to move on and launch a new long-term project: the ZEN*comp*. The new corpus is meant to complement the ZEN and will consist of a selection of periodical news publications again gathered in ten-year intervals. Starting in 1621, the ZEN*comp* will add substantial material to the ZEN, including early corantos, newsbooks, and news-pamphlets.

The ZEN project is supervised by Udo Fries. Peter Schneider co-ordinates the technical support and publication of the ZEN, while Patrick Studer is in charge of preparing and finalising the ZEN corpus for publication.

The ZEN team is happy to welcome Nicholas Brownlees, of Florence University, author of the recently published book *Corantos and Newsbooks: Language and Discourse in the First English Newspapers (1620-1641)*. As a further development to his research he has agreed to work on the ZEN*comp*.

Udo Fries: ufries@es.unizh.ch

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

The transcription of letters written to Richard Orford in the Leghs of Lyme collection will actually occupy us for the remainder of this second and final year of the present research funding. We have erred on the side of inclusiveness. This means, however, that the material is not uniform, nor is it balanced in its coverage, whether by kind of writer or by topic. For example, there is a good deal of plain 'business English' of the late eighteenth century; there are some letters which represent more dialectally marked and uneducated writers; there are

some from writers higher up the social scale. Nevertheless there is a lot of good linguistic (and socio-historical) material from local and relatively unselfconscious writers. Here, for instance, is part of a letter from James Grimshaw to Orford, written in 1779 (crossings-out silently omitted for this report):

> There has not been such Mobing this many years as at this time there is a Mob at this present time that is puling down all the Carding and Spining Machines that go by water, three is allready pul'd Down

Or in 1790:

> I have sent about M^rs^.. Hancocks Boxes and there is no boxes come for her. – Molley sends 12 bottles of Vinegar. – We have this Day had the vilantest storm of rain and hale I ever seed; attended with Lightening and Thunder the Lightning killed us two Cows Just by the corner of the Garden; which I leave to you to aquant M^r^. Legh of.

The project is directed by David Denison in collaboration with Linda van Bergen and (in 1999–2000) Joana Proud. With the aid of a renewed bursary from the John Rylands Research Institute, Linda van Bergen and Joana Proud are working part-time during 1999–2000 to select, transcribe and annotate the Orford documents. So far over 200,000 words have been transcribed and proofread. Although we hope to be able to cover all of the Orford materials, time must be left for putting the corpus into a usable form. Suggestions as to suitable search and display software will be welcome, as will comments on the wisdom (or otherwise) of attempting to create a parallel normalised text.

David Denison: d.denison@man.ac.uk

CONCE – A Corpus of Nineteenth-century English

For information on the CONCE corpus, see the present publication, pp 85–98.

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

HISTORICAL DICTIONARIES, GRAMMARS AND ATLASES

Dictionary of Old English Project

The Dictionary of Old English Project is pleased to announce the publication of the first CD-ROM of its Corpus in January 2000. The CD-ROM Corpus is the most up-to-date and accurate version of the Corpus available. In preparation for CD-ROM publication, we undertook a major revision of the Corpus: correcting errors, updating editions, regularizing layout, searching for and correcting anomalies. The Corpus is in Standard Generalized Markup Language (SGML), and conforms fully to the guidelines of the Text Encoding Initiative. We also provide an HTML rending of the SGML version, so that users can point their Web browsers to it for ease of viewing. The Dictionary of Old English Corpus on CD-ROM is available from the project for $200 US (contact: corpus@doe.utoronto.ca). We also distribute the Corpus on UNIX tar(1) tapes. With the publication of the Corpus on CD-ROM we have discontinued issuing it on diskette.

Work is continuing on the tagging of the legacy data of the *Dictionary* (the fascicles already published) in preparation for CD-ROM publication of *A* through *F*. As of 15 February 2000, five letters (*A, Æ, C, D*, and *E*) of the six published fascicles of the *Dictionary* have been tagged. First-stage tagging of *B* is currently underway. Once *B* is tagged, the markup of all the legacy fascicles will have been completed, and only *F* remains to be done.

On the editorial side, work is progressing on completing entries for the letter *F*, and strong advances have been made on the letters *G* and *H*. In addition to writing, revising, and inputting new entries, we have begun as time allows, the systematic revision of previous fascicles in preparation for their inclusion with *F* on the first CD-ROM of the *Dictionary.* Revisions encompass the correction of substantive errors, the updating of OE texts and Latin source material, and the replacement of hand-inserted special characters by electronically generated ones now available through more sophisticated software. Our most frequent implementations involving special characters have been the tagging and printing of Greek and diacritics for our earliest fascicles and the tagging and printing of crossed *l*s and crossed *thorns* throughout. We are delighted to report that *B*, the largest of the fascicles published so far, has been comprehensively corrected and updated, together with *E*. Good progress has been made on fascicle *C*, one of our earliest letters, and therefore in greater need of revision. As soon as *C* is completed, we will proceed through *A, Æ* and *D*. Our goal is to maintain and update our electronic files on a regular basis as we continue our push through the alphabet. We believe this will be an advance over earlier dictionaries which,

lacking the technological developments available today, were forced to issue supplements to update their work. Our wish is to achieve a good balance between writing the rest of the *Dictionary* and maintaining updated files of past work. Repair work of this kind is absolutely crucial to prevent the databases we create from becoming fossils.

We continue to be pleased at the response of our users to the searches they can conduct on the *Dictionary of Old English Corpus on the World-Wide Web*. This past year more than fifty institutions throughout the world acquired site licenses. Information on the Web Corpus is available from the project's webpage: http://www.chass.utoronto.ca/oec/

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

The Thesaurus of Old English: Future plans

A Thesaurus of Old English, or *TOE*, edited by Jane Roberts and Christian Kay with Lynne Grundy (King's College London Medieval Studies XI) came out in 1995. This pilot study for the forthcoming 'Historical Thesaurus of English' is now sold out. A second edition is planned with Rodopi. The editors are also exploring the possibility of generating a dictionary from the *TOE* database.

Jane Roberts: jane.roberts@kcl.ac.uk

The Historical Thesaurus of English

The *Historical Thesaurus of English* project entered the new century with about three years work left to do. Approximately 90 per cent of the material has been edited, and over 80 per cent is held in an Ingres database, which now contains some 545,000 records. The largest section so far is the ANIMAL KINGDOM, with 42,571 records, edited by Cerwyss O'Hare. Work is currently proceeding at Glasgow on editing the large fields of SPACE, MENTAL ACTIVITY, and PLANTS, while Angus Somerville of Brock University, Canada, is completing the classification of LANGUAGE. Thereafter, only two major sections remain to be edited: ACTIVITY and EXISTENCE, both with around 50,000 words.

As with any large project, there also remain a few sections which need revision or repositioning, and entries to insert or correct. We are fortunate in having Professor M.L. Samuels, who founded the project in 1965, as our principal

proof-reader. Data collection from our main source, the *Oxford English Dictionary* (*OED*), is still underway, but although we want our data to be as comprehensive as possible, we have to stop somewhere, and have decided that the cut-off point will be the three *Additions* volumes to the *OED*.

The main database is held on a Sun Ultra 10 workstation in Access and in OpenIngres 2.01, with Hummingbird I Query (the newest version of what was formerly called GQL) as interface. We have our own computing officer, Flora Edmonds, who works on the development of the database in conjunction with Irené Wotherspoon, Senior Research Assistant.

The main purpose of the *Historical Thesaurus* is to provide a new kind of source material for the study of the English vocabulary. Its underlying assumption is that new insights can be gained into many aspects of the history of a language by examining words in their semantic context. To the best of our knowledge, no such resource exists in either paper or electronic form for any other of the world's languages. Data from the *OED* and from *A Thesaurus of Old English* (Roberts and Kay, 1995) are presented in ordered semantic categories, giving the scholar access to words used to express concepts at any period from Old English to the present. Each word is uniquely retrievable from the database, which will be used both to generate a printed text and as an electronic resource, both to be published by Oxford University Press. Web publication of the *Historical Thesaurus* alongside the *OED* will eventually create a uniquely powerful research tool, enabling scholars to view the wealth of information in the *OED* within a semantic framework. Samples of data from several fields can be viewed on our website, http://www.arts.gla.ac.uk/EngLang/thesaur/homepage.htm.

In addition to those mentioned above, the *Historical Thesaurus* employs three part-time research assistants, a part-time computing officer and five student assistants. Funding has come from the British Academy, the Carnegie Trust for the Universities of Scotland, the Leverhulme Trust and the University of Glasgow. The project operates under the aegis of the Glasgow Institute for the Historical Study of Language along with the *Middle English Grammar Project* (Jeremy Smith and Simon Horobin) and the *Anglo-Saxon Plant-Names Survey* (Carole Biggam). We will be represented at the Association for Literary and Linguistic Computing Conference, to be held in Glasgow from July 21–5, 2000, and at the Thirteenth International Conference on English Historical Linguistics in Spain in September 2000.

Christian J. Kay: c.kay@englang.arts.gla.ac.uk

The Middle English Grammar Project

I. Background

Since the publication of the *Linguistic Atlas of Late Mediaeval English* [LALME][1] and other important surveys of Middle English [ME] dialectology, such as Kristensson's *Survey of Middle English Dialects*[2], previous ME Grammars, notably Jordan's *Handbuch der mittelenglischen Grammatik: Lautlehre*[3], have become seriously outdated. The Middle English Grammar Project[4] aims to provide scholars with an authoritative reference-point for every level of language during the Middle English period: transmission (handwriting, spelling and phonology), grammar and lexicology. The project is focused on the Department of English Language at Glasgow University, in collaboration with Stavanger College, Norway, and it has close links with the Institute for Historical Dialectology at Edinburgh University. The various project components will be addressed in the following order: transmission, lexicology, grammar. Work on the spelling component is now underway. A central focus of this work is the creation of an electronic corpus which will be described in this report.

II. The corpus consists of two main elements:

1. Transcriptions

Electronic transcripts of 3,000 word tranches are currently being assembled of all the texts listed as localised or localisable in LALME. These will be supplemented with similar tranches of texts localised by the Linguistic Atlas of Early Middle English [LAEME]. These transcriptions have been prepared using the same conventions as adopted for LAEME and are described in full elsewhere.[5]

II. Database

The transcribed material is subsequently classified and entered into a database. The project has developed a new classificatory system, on the model of the 'standard lexical sets' used by J. Wells in *Accents of English* (1982)[6], to group words whose behaviour in Present-Day English [PDE] is similar (eg the STRUT group, the FLEECE group etc). This classificatory model, termed 'standard orthographic sets', groups words according to correspondences in their PDE spelling. A 'standard orthographic set' is therefore the set of forms which now appear with (eg) PDE <ea>, PDE <sh> etc. The database also includes the base form for each of the words entered, thereby permitting reference to the traditional method of classifying ME forms, according to the reflexes of Late West Saxon [LWS] etc. Individual entries also comprise the manuscript shelfmark,

the modern title of the work and author, and the LALME Linguistic Profile number plus localisation and grid reference. In addition to these, data information is included describing the genre of the text and its manuscript context, and each entry in the database contains details concerning the production and early history of the manuscript. A sample record from the database is as follows:

Group:	U
Sub Group:	u
PDE Spelling:	church
ME Spelling:	CHERECHE
Frequency:	1
MS Reference:	Laud Misc 471
Text:	Kentish Sermons
LALME Reference:	LP 6050
Source:	Microfilm
County:	Kent
Grid Reference:	562 174
Date:	13a
Genre:	Religious
Hand:	Anglicana
Rhyme:	N/A
Base Form:	CIRICE
Language:	OE

The database may be manipulated in a variety of ways. For example, a simple search might extract all ME spellings of words containing PDE 'CH', or all ME spellings of PDE 'church', or all occurrences of the ME spelling 'chereche'. Alternatively one might wish to examine the reflexes of WS <y>, or <sc> within the corpus. More complex searches allow the manipulation of this same information according to particular counties, texts, authors, or manuscripts. Thus one might wish to retrieve all spellings of 'such' in London manuscripts of *Piers Plowman* or all occurrences of a particular spelling in rhyming position in Chaucer's poetry. The number and range of possible searches are very great, and such information will be of use to scholars working in a variety of related areas as well as for the immediate purposes of the project. The results of our survey of ME spelling practices will be published in conventional book form, although it is also envisaged that the database will be made available by publication on CD-ROM or via the Internet.

References

1. Samuels M.L., A. McIntosh and M. Benskin (eds), *A Linguistic Atlas of Late Mediaeval English*. [LALME] Aberdeen: Aberdeen University Press, 1986.
2. Kristensson, G., *A Survey of Middle English Dialects*. Lund: Gleerup, 1967 etc.
3. Jordan, R., *Handbuch der mittelenglischen Grammatik: Lautlehre*. trans. E.J. Crook. The Hague: Mouton, 1974.
4. The Middle English Grammar Project is funded by the British Academy.
5. Horobin, S.C.P. and J.J. Smith, 'A Database of Middle English Spelling', *Literary and Linguistic Computing* (forthcoming).
6. Wells, J. C., *Accents of English*. 3 vols. Cambridge: Cambridge University Press, 1982.

Simon Horobin: s.horobin@englang.arts.gla.ac.uk
Jeremy Smith: j.smith@englang.arts.gla.ac.uk

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

The work of transcribing and tagging early Middle English texts to enlarge the corpus has continued with the help of a year's funding from the British Academy. The project for the year has been to complete the tagging and linguistic analysis of the seven early ME manuscripts containing Ancrene Riwle and the Katherine Group of texts, all of which originate in the West Midlands. Since the Ancrene Riwle is very long, its language in each version will be represented not by the whole text but by a sample of about 15,000 words.

Since last year's report, the tagging of the sample from BL Cotton Nero A.xiv has been completed, and those from Cambridge, Corpus Christi College 402 and BL Cotton Cleopatra C.vi, scribe A (main scribe) have also been done. This has added a further 39,564 words to the corpus of tagged texts. The transcribing and tagging of the corrections to the Cleopatra text by scribe B (perhaps the author of Ancrene Riwle himself), and of the version in Cambridge, Gonville and Caius MS 234/120 are under way.

Where possible, the dialects of all these texts will be localised, mapped and related to other important, much-studied texts from the same region, such as the

work of the Worcester Tremulous Scribe and the four 13th-century verse miscellanies (Oxford, Bodleian Library, Digby 86; Oxford, Jesus College MS 29; BL Cotton Cleopatra A.ix and Cambridge, Trinity College B.14.39 (323)). The West Midlands is the only area for which the density of coverage of texts from the early Middle English period is comparable to that presented in *A Linguistic Atlas of Late Mediaeval English*, so this should make a significant contribution towards the historical linguistic geography of the area.

As the corpus of tagged texts grows larger (now well over 400,000 words), a certain amount of time and effort has to be put into 'routine maintenance': error spotting and making sure that the tags are consistent across the corpus. This year I have also undertaken a major revision of the tags that indicate verb negation in order to make them more sensitive to syntactical variation. Our work at the IHD is primarily geared towards the making of linguistic atlases, and our system of tags therefore reflects a bias towards lexical and morphological variation. But it is possible to adapt the lexico-grammatical tags to take account of some syntactical elements also. A preliminary study towards a corpus-based approach to aspects of early Middle English negation is in progress.

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