English historical corpora: Report on developments in 1995

Merja Kytö and Matti Rissanen
University of Helsinki

At the First International Colloquium on English Diachronic Corpora, held in March 1993 at St Catharine's College, Cambridge, it was decided that the English Department of the University of Helsinki would be responsible for collecting and distributing information in the fields covered by the Colloquium. Consequently, Merja Kytö and Matti Rissanen have chaired short historical corpus workshop sessions at the ICAME Conferences held in Zurich in 1993 and in Aarhus in 1994. A more extensive two-day workshop was held in connection with the Toronto ICAME meeting in 1995. The next two-day workshop will take place in Helsinki and Stockholm, and on board the ferry between the two cities, immediately preceding the Stockholm ICAME Conference, in May 1996.

The first written report on the developments in English historical corpora, thesauruses, atlases and dictionaries was published in *ICAME Journal* 19, 1995, pp. 145–158. The papers and reports of the Toronto workshop will be published as a Conference Volume.

The present report will supplement the reports included in *Corpora Across the Centuries: Proceedings of the First International Colloquium on English Diachronic Corpora* (Amsterdam & Atlanta, GA: Rodopi, 1994) and the report published in *ICAME Journal* 19. References to those reports are given in brackets after each entry.

We are indebted to the scholars working on corpus studies and methodology for sending us the news for this report.

Matti Rissanen
mrissanen@cc.helsinki.fi

Merja Kytö
mkyto@cc.helsinki.fi
merja.kyto@engelska.uu.se

CORPORA AND DATABASES COMPLETED

1. Michigan Early Modern English Materials

Keyboarded with support from the National Endowment for the Humanities from 1968 to 1974, Michigan Early Modern English Materials was the first completed database for a historical period of the language. It consists of 50,000 citation slips collected for the Oxford English Dictionary and the successor project, the Early Modern English Dictionary (never completed). Each contains about 50 words of English from the period 1475 to 1700. The majority are the slips selected to illustrate nuances of the English modal verbs (and *have* and *be*) in the belief that this collection would constitute a random sample from all texts read for the two dictionaries. Additional slips from the antedating collection and for words of intermediate frequency were also encoded and added to the materials. The entire file is 16 megabytes and can be freely downloaded. However scholars can have unrestricted access through the World Wide Web. It contains its own search engine and will accommodate most searches possible with other corpora. Abbreviated titles generally follow those in the OED; specific details are found in the handbook (MICHIGAN EARLY MODERN ENGLISH MATERIALS, Richard W. Bailey and others, 1975), issued when a portion of the corpus was published on microfiche. The address of the corpus is: http://www.hti.umich.edu/dict/memem/

Richard W. Bailey: rwbailey@umich.edu

2. 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

3. The Helsinki Corpus of Older Scots (1450–1700)

The Helsinki Corpus of Older Scots has been compiled as a supplement to the Helsinki Corpus of English Texts: Diachronic and Dialectal. It contains text extracts from early prints or the best editions available. The full text of shorter works has been included; in longer texts, the sample size varies from 5,000 to 40,000 words of running text. The majority of the texts originate from the Central Scots area (East Mid Scots); a small number of texts represent North-East Scots.

The Corpus of Older Scots offers an opportunity to consult a much wider range of texts representing the variety than has previously been available to researchers. It contains approximately 850,000 words of running text; the periodization and number of words per period are as follows:

1450–1500	95,900
1500–1570	200,500
1570–1640	302,300
1640–1700	247,000
Total	845,700

The texts represent fifteen different prose genres: acts of Parliament, burgh records, trial proceedings, histories, biographies, travelogues, diaries, pamphlets, educational treatises, scientific treatises, handbooks, private letters, official letters, sermons and the Bible. A number of these mainly non-literary genres were assumed to reflect spoken language or language used in informal settings.

The Corpus contains information on sociohistorically relevant extralinguistic variables such as the author's rank, age and sex, description of audience, level of formality, relationship to spoken language, interaction, text category, whether the text was printed or not, and, in private and official letters, participant relationship.

For further information on the Corpus, please contact:
Dr Anneli Meurman-Solin
Department of English
P.B. 4 (Yliopistonkatu 3)
FIN-00014 University of Helsinki, Finland
E-mail: Anneli.Meurman-Solin@Helsinki.FI
Fax: +358-0-19123072

Permission to use the material must be obtained from the Department of English, University of Helsinki. Please write to:
Professor Matti Rissanen
Department of English
University of Helsinki
P.O. Box 4 (Yliopistonkatu 3)
FIN-00014 University of Helsinki, Finland

The Helsinki Corpus of Older Scots is available from:
Norwegian Computing Centre for the Humanities
Harald Hårfagresgate 31
N-5007 Bergen
Norway
E-mail: icame@hd.uib.no
WWW: http://www.hd.uib.no/icame.html
Fax: +47 55 58 94 70; tel: +47 55 58 2954/5/6

Oxford Text Archive
Oxford University Computing Service
13 Banbury Road
Oxford OX2 6NN
United Kingdom
E-mail: archive@vax.oxford.ac.uk
WWW: http://info.ox.ac.uk/~archive
Fax: +44-865-273275

(*Corpora Across the Centuries*, pp. 53–63)
(*ICAME Journal* 19: 149; 157–8)
Anneli Meurman-Solin: anneli.meurman-solin@helsinki.FI

4. The Corpus of Late Modern English Prose

The Corpus of Late Modern English Prose, compiled in Manchester by David Denison has been heavily used in the writing of the 'Syntax' chapter of the *Cambridge History of the English Language*, IV, *1776–Present Day*. It has also been requested for linguistic investigation by a number of colleagues in Europe and North America.

(*Corpora Across the Centuries*, pp. 7–16)
(*ICAME Journal* 19: 146)
David Denison: d.denison@man.ac.uk

NEW CORPUS PROJECTS

5. A Corpus of Dialogues (1550–1750)

Jonathan Culpeper (Lancaster University) and Merja Kytö (Uppsala University) are collaborating on a project aiming at a corpus of texts reflecting spoken dialogue from 1550 to 1750. The only way of recovering the speech of the past is to turn to written documents. One can find brief extracts of speech-based texts in, for example, the Helsinki Corpus. Since change is mostly actuated in spoken language, a large computerised corpus offering easy access to a structured and contextualised selection

of texts would be a valuable resource. Dialogues are of particular interest as they constitute data for diachronic studies in interactive communication.

Our objective is to construct a corpus of a good million words, comprising texts varying from supposedly recorded speech events, such as some trial proceedings and parliamentary journals, to constructed imaginary dialogue found in, for example, drama and prose fiction. Preliminary planning was undertaken in October 1995, and with the help of a grant from the British Academy, work on the project will commence in June 1996.

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

6. A Corpus of 19th Century Texts

As part of a project aiming at a corpus of 18th and 19th century texts, Merja Kytö (Uppsala University) and Juhani Rudanko (Tampere University) are currently collaborating on a collection of texts representative of 19th century writing.

To enable comparisons with e.g. the Helsinki Corpus of English Texts, the corpus will focus on central text types such as fiction, science, history writing, drama, handbooks and correspondence. The subperiodization will make it possible to study texts drawn from the beginning, middle and end of the century.

The final version of the corpus (1700–1900) will consist of a good million words. The work was commenced in autumn 1995.

Merja Kytö: merja.kyto@engelska.uu.se
Juhani Rudanko: f1juru@uta.fi

7. Crunching from Old English to Jane Austen

A new corpus activity in the Department of English at Manchester (carried out by David Denison and/or Linda van Bergen) has been to prepare a number of texts available from the OTA or ICAME for use with WordCruncher for DOS 4.5, marking them up for the proper display of special characters using DOS codepage 850, indexing them in some cases for sentence boundaries, for the convenient display and retrieval of author, addressee, text, page, and so on. Texts indexed include the

Letters of Jane Austen, vol. 1 of the *Paston Letters*, the Middle English portion of the Helsinki Corpus, and the Toronto Corpus of Old English. These are contributing to various kinds of linguistic and literary research in the Department. More texts will be added soon, perhaps using WordCruncher for Windows, and probably involving partial syntactic tagging.

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

PROGRESS OF EARLIER PROJECTS

8. The ARCHER Corpus

The ARCHER Corpus, in preparation under the supervision of Douglas Biber (Northern Arizona University) and Edward Finegan (University of Southern California), aims at a c. 1.7 million-word collection of English texts representative of written and speech-based registers from 1650 to the present. The texts have now been stored in magnetic format, but copyright settlements have proved more complicated and time-consuming than expected. We hope, however, that a preliminary version of the corpus will be available (in ASCII format) from the Norwegian Computing Centre for the Humanities (Bergen) by the end of 1996.

(*Corpora Across the Centuries*, pp. 3–6)
(*ICAME Journal* 19: 148)
Douglas Biber: douglas.biber@nau.edu
Edward Finegan: finegan@mizar.usc.edu

9. ICAMET – Innsbruck Computer Archive of Middle English Texts

The Innsbruck Corpus, which is a full-text data base of selected Middle English prose, comprises about a hundred files. At present, these are still being correlated as to special characters, format and layout. The first samples will be available on diskette for individual users after September 1996. In 1997, a CD-ROM of the whole corpus is going to be produced.

For the time being, the availability has to be limited to files which are free of copyright restrictions. As far as the other files are concerned, we are still waiting for the EETS to give us licence.

The files will come in two different versions: in a DOS version, which is well-legible and reliably presented in WORD 5 and Norton Commander, but uses some higher ASCII characters; and in a lower ASCII version with various coded characters, which are, of course, no problem to the machine, but make it difficult for a human reader to scan the text.

As an appendix to ICAMET, 254 English letters, dated from 1386 to 1689 and taken from various collections, will also be available on three separate diskettes as of September 1996, likewise a first version of the manual of the whole corpus. The parameters of the various files, planned as a 'header' for each file, are under work and will not be available before 1997.

(*Corpora Across the Centuries*, pp. 41–52)
(*ICAME Journal* 19: 150)
Manfred Markus: manfred.markus@uibk.ac.at

10. The Corpus of Early English Correspondence (CEEC)

Our project on historical sociolinguistics was launched at the Department of English of Helsinki University in 1993. For the sociolinguistic purposes of the project, a socially representative corpus of early letters has been compiled in the course of the last two and a half years. This 2.4 million-word corpus, called 'The Corpus of Early English Correspondence', consists of letters written by 677 individuals, covering the period from 1420 to 1680. The key variables considered in the selection process included the writer's socioeconomic status, gender, age, provenance, relation to the recipient as well as social and geographical mobility.

At present, the corpus is being checked and annotated to include such letter-specific information as authenticity, time of writing and the writer's relation to the recipient. A separate database has been created on the social backgrounds of the writers. At the completion of the project, provided that the necessary permissions will be granted by the copyright holders, the corpus will be made available to the research community in electronic form.

Reference:

Nevalainen, T. & H. Raumolin-Brunberg, eds. (1996). *Sociolinguistics and language history: Studies based on The Corpus of Early English Correspondence*. Amsterdam: Rodopi.

(*ICAME Journal* 19: 147)
Terttu Nevalainen: tnevalainen@cc.helsinki.fi
Helena Raumolin-Brunberg: raumolinbrun@cc.helsinki.fi

11. ZEN – The Zurich English Newspaper Corpus

The Zurich English Newspaper Corpus is a collection of texts taken from English (London) newspapers from the mid-1660s to the end of the 18th century.

By the end of 1995 about 1,000,000 words have been keyed in; the texts chosen so far appeared between 1671 and 1791. These texts have been collected in 10-year intervals. In its final shape the corpus will be divided into four 20-year periods (1671–1691, 1701–1721, 1731–1751, 1761–1781) with additional sections covering the decades before and after 1671 and 1781, respectively.

The largest part of our collection are texts from *The London Gazette* – the only paper which appeared throughout the period investigated. But the Corpus will also include a fair selection of most of the other London newspapers of the 18th century, many of which were, however, only very short-lived. We will try to achieve a fair balance between the individual newspapers.

We have begun assigning labels to individual text classes (e.g. <foreign news>, <home news>, <proclamation>, <ship news>, <crime>, <births>, <deaths>, <advertisement>, <address>, or <letters>). In many instances there are overlaps and frequently decisions taken by the compilers are problematic. In addition, it has become apparent that some sections of the corpus are underrepresented, while others (e.g. advertisements) are overrepresented.

A major problem surfacing at this stage goes back to the history of the compilation of the corpus. The majority of texts were keyed in by undergraduate students. The number of errors and other idiosyncracies is high and we are trying to eliminate as many of them as possible. This is the major cause for delays in the publication of a preliminary version.

We have not finally decided in which way we will make the corpus available. We are investigating several options that take into consideration that ours will basically be an open-ended corpus. There may also be a tagged or semi-tagged version of the corpus.

A paper on ‘The Vocabulary of ZEN – Implications for the Compilation of a Corpus’ will be published in the proceedings of the ICAME 1995 conference. It gives more information on the corpus.

(*Corpora Across the Centuries*, pp. 17–18)
(*ICAME Journal* 19: 148–9)
Udo Fries: ufries@es.unizh.ch

HISTORICAL ATLASES AND CONCORDANCES

12. Progress on the Historical Thesaurus of English 1995

The Historical Thesaurus of English project continues to make progress. The team was much heartened by the publication in 1995 of *A Thesaurus of Old English* (Jane Roberts and Christian Kay with Lynne Grundy, King's College London Medieval Studies XI, 1995, 2 vols., xxxv + 1555, ISBN 0 952211904). In addition to being a research tool in its own right, *A Thesaurus of Old English* formed a pilot study for the Historical Thesaurus.

Continued funding from the Leverhulme Trust and the British Academy as well as support from the University of Glasgow has enabled us to increase staffing levels. Two of the three major sections, the External World and Society, are almost complete, and a start has been made on the more abstract lexis of the third section, Mind. Additions to the database in 1995 included Esteem, Possession, Law and Society, while work is in progress on Power, Mind and the huge section on Animals, currently the largest single category.

Computing facilities have been upgraded and over 60% of the data, comprising 430,000 records, has now been typed into the database by employment trainees. Work continues on developing the database structure in Ingres, with Access being explored as a possible front end.

(*Corpora Across the Centuries* 111–20, 155–61)
(*ICAME Journal* 19: 152–3)
Christian Kay: cjkay@human.gla.ac.uk

13. The Linguistic Atlas of Early Middle English, Institute for Historical Dialectology, University of Edinburgh

The Corpus of Early Middle English Texts

Transcribing and tagging of early Middle English texts has continued steadily over the last year. The corpus of early Middle English texts transcribed and tagged now consists of 161 texts from 53 different manuscripts, of which 37 (from 30 different manuscripts) have been added since the last report. (See the updated list below; new texts are marked with a preceding *.) In addition, the tagging of both manuscript copies of The Owl and the Nightingale (for each of which only 300 lines had been tagged by the end of 1994) has now been completed (1794 lines in each manuscript). To date about 200,000 words of text have been tagged, about 70,000 since January 1995.

From the tagged corpus, dictionaries are generated which so far contain 20,169 different tags describing 29,815 different forms. The tagged corpus now represents 67 different hands or types of early ME language, of which 54 have been given provisional placings on the map. Working maps of 21 different items have been generated from the files of tagged texts, using a mapping program devised by Dr Keith Williamson. We have recently obtained a more advanced software mapping package which will enable us to produce different kinds of maps to make analysis easier and presentation more flexible. These working maps will prove increasingly useful in the fitting of new Text Profiles.

Applications of the Project Methods in Middle English Studies

A great deal of time in 1995 was spent sorting out, by means of tagging and linguistic comparison, the complex textual history of the Titus MS of Ancrene Riwle and the Katherine Group. The resulting paper illustrates the power of the methodology we are using in the project to highlight textual and stemmatological, as well as linguistic, relationships between manuscripts and texts. The investigation revealed that a single scribe was responsible for writing six somewhat differing types of early ME. Two of these proved internally consistent enough to be given locations on the dialect map. Although such detailed studies take time, it is proving necessary to subject many of the more linguistically complex early ME texts to this kind of detailed scrutiny before it is possible to say where their language(s) belong(s). See Margaret Laing and Angus McIntosh, 'The Ancrene Riwle, the Katherine Group Texts and þe Wohunge of ure Lauerd in BL Cotton Titus D xviii' *Neuphilologische Mitteilungen* 96 (1995), 235–263 and cf. Margaret Laing and Angus McIntosh, 'Cambridge, Trinity College MS 335: Its Texts and Their Transmission' in *New Science out of Old Books: Studies in Honour of*

A.I. Doyle, ed. Richard Beadle and A. J. Piper (Aldershot: Scolar Press, 1995), pp. 14–52.

List of Tagged Texts in the Early Middle English Corpus to Date

(* indicates that the text has been added since the last ICAME report)

*Cambridge, Corpus Christi College 444 fols. 1r–81r: Genesis and Exodus (only 1r–41r tagged)
*Cambridge, Emmanuel College 27, fols. 111v, 163r–163r: lyrics
Cambridge, Fitzwilliam Museum, McClean 123, fols. 115r–120r: Poema Morale
Cambridge, St John's College A.15, fols. 72r; 120v: lyrics including Candet Nudatum Pectus
*Cambridge, St John's College F.8: 17 fragments of the Proverbs of Hending
Cambridge, Sidney Sussex 97 (D.5.12), fol. 111r: Candet Nudatum Pectus
*Cambridge, Trinity College 43 (B.1.45), fols. 24r–v, 41v–42r: verses and two sermons
Cambridge, Trinity College 335 (B.14.52), (a) fols. 2r–9v: Poema Morale; (b) pp. 1–157: Trinity Homilies (34 Homilies, three hands)
Cambridge University Library Ff.II.33, fols. 20r–v, 22r–24r, 27v–28r, 45r–47r, 48r–50r: 48 Documents from Bury St Edmunds, Suffolk
*Cambridge University Library Ff.VI.15, fol. 21r: Ten Commandments
*Carlisle, Cumbria Record Office, D/Lons/L Medieval deeds C1: Gospatric's Writ
Durham, Dean & Chapter Library A III 12, fol. 49r: the lyric Candet Nudatum Pectus
Herefordshire Record Office AL 19/2, Registrum Ricardi de Swinfield fol. 152r: Bromfield Writ
London, British Library, Additional 11579, fols 35v–36v; 72v–73r: lyrics including Candet Nudatum Pectus
*London, British Library, Additional 15350, fols. 115v–117r: Vision of Edwin
*London, British Library, Additional 23986, verso of roll: Interludium de Clerico et Puella
*London, British Library, Arundel 248, fols. 154r–155r: four lyrics
*London, British Library, Arundel 292, fols. 4r–10v: The Bestiary

London, British Library, Cotton Caligula A ix, (a) fols. 233r–239v line 13; 240r line 6 –241v line 15: The Owl and the Nightingale, language 1; *(b) fols. 239v line 14 – 240r line 5; 241v line 16 – 246r: The Owl and the Nightingale, language 2

*London, British Library, Cotton Cleopatra C vi, hand D, (a) fols. 22v–23r, 57v: verses and a sermon; (b) additions and corrections throughout

*London, British Library, Cotton Galba E ii, fols. 30r–v: documents from Benet Holme

*London, British Library, Cotton Otho B xiv, fol. 263r: fragment of a Ramsey Register

*London, British Library, Cotton Roll ii 11, (a) language 1: 3 documents from Crediton; (b) language 2: document from Crediton

London, British Library, Cotton Titus D xviii, (a) fols. 14r–40r: Ancrene Riwle (part of language 1); (b) fols. 40ra1–40vb6, 52va17–55ra25, 56va7–61rb22, 67rb17–68ra2, 69ra2–70ra1: Ancrene Riwle (language 2); (c) fols. 105v–112v: Sawles Warde; *(d) fols. 112v–127r: Hali Meidhad (e) fols. 127r–133r: Wohunge of ure Lauerd; *(f) fols. 133v–147v: Seint Katherine

London, British Library, Egerton 613, (a) fols. 7r–12v : Poema Morale (E); (b) fols. 64r–70v: Poema Morale (e)

*London, British Library, Harley 978, fol. 11v: Svmer is icumen in

London, British Library, Stowe 34, pp. 1–95: Vices and Virtues, hand A, pp. 1–74 line 17 and 74 line 22 –75 line 3; hand B, pp. 74 lines 17–22 and 75 line 3–95.

*London, Corporation of London Record Office, Liber de antiquis Legibus, fols. 160v–161v: Prisoner's Prayer

London, Lambeth Palace Library 487, fols. 59v–65r: Poema Morale

*London, Lambeth Palace Library 499, fols. 64v–69r, 125v: alliterative lyrics

*London, Lincoln's Inn Hale 135, fol. 137v: Nou sprinkes the sprai

*London, PRO, E 164/28, Register of Ramsey Abbey, (a) fols. 52v–53r, 59v–60r, 159v–170v: hand A; (b) fol. 229v: hand B

London, PRO, Patent Rolls 43 Henry III, m. 15.40: Huntingdon copy (enrolled) of the Proclamation of Henry III of 18 October 1253

Maidstone Museum A.13, (a) fol. 93r: Proverbs of Alfred; (b) fol. 93v: Death's Wither-Clench; (c) fol. 243v: Three Sorrowful Things

*Oxford, Bodleian Library, Additional E.6 roll of 4 membranes, (a) hand A: Sayings of St Bernard; (b) hand B: XV Signs before Doomsday and Pater Noster

*Oxford, Bodleian Library, Bodley 26, fols. 107r–108r: macaronic sermon
Oxford, Bodleian Library, Bodley 34, (a) fols. 52r–71v: Hali Meidhad; (b) fols. 72r–80v: Sawles Warde
Oxford, Bodleian Library, Bodley 42, fol. 250r: lyrics including Candet Nudatum Pectus
*Oxford, Bodleian Library, Bodley 652, fols. 1r–10v: Iacob and Iosep
Oxford, Bodleian Library, Digby 4, fols. 97r–110v: Poema Morale
Oxford, Bodleian Library, Digby 45, fol. 25r: the lyric Candet Nudatum Pectus
Oxford, Bodleian Library, Digby 55, fol. 49r: lyrics including Candet Nudatum Pectus
*Oxford, Bodleian Library, Hatton 26, fol. 211r: Ten Commandments and Seven Gifts
*Oxford, Bodleian Library, Laud Misc. 471, (a) fol. 65r: Death's Wither Clench; (b) fols. 128v–133v: Kentish Sermons
*Oxford, Bodleian Library, Laud Misc. 636, fols. 88v–91v: Peterborough Chronicle, final continuation
Oxford, Bodleian Library, Rawlinson C 317, fol. 89v: the lyric Candet Nudatum Pectus
*Oxford, Bodleian Library, Rawlinson C 510, fol. 3r: fragment of a lyric
Oxford, Bodleian Library, Tanner 169*, p. 175: Stabat iuxta crucem Christi
Oxford City Archives, Town Hall, St Aldates, H 29: Oxford copy of the Proclamation of Henry III of 18 October 1253
Oxford, Jesus College 29, (a) fols. 156r–168v: The Owl and the Nightingale; (b) fols.169r–174v: Poema Morale; *(c) fols. 187r–188v: A Luue Ron
Stratford-upon-Avon, Shakespeare Birthplace Library, DR 10/1408, Gregory Leger-Book, pp. 23–24: Coventry Writ
*Wells Cathedral Library, Liber Albus I, (a) fol. 14r: four documents from Wells, language 1; (b) fols. 17v–18r: five documents from Wells, language 2

(*Corpora Across the Centuries*, pp. 121–141)
(*ICAME Journal* 19: 154–5)
Margaret Laing: esss09@castle.ed.ac.uk

14. Concording Early Thirteenth-Century 'AB' Language (Ancrene Wisse and the 'Katherine' Group Texts)

The corpus for this project consists of the text of Ancrene Wisse as written by the principal scribe of Cambridge, MS Corpus Christi College 402 ('A' of so-called 'AB' language) and the texts of the 'Katherine Group' from Oxford, Bodleian Library, MS Bodley 34 ('B' of 'AB' language): manuscript texts of the associated 'Wooing Group' are also included. The texts are held in word-processed and electronic form, lineated in accordance with the manuscript lineation, and scribal word-division is edited to provide interrogatable morphemes for concordancing programmes.

In 1993, this project published *A Concordance to Ancrene Wisse: MS. Corpus Christi College, Cambridge, 402* edited by Jennifer Potts, Lorna Stevenson and Jocelyn Wogan-Browne (Cambridge: D.S. Brewer, 1993), 1249 pp, ISBN 0 85991 395 3. This concords vernacular forms, gives sample listings and indices for high frequency forms and provides alphabetical lists of vernacular and other forms, indexes of proper names, a descending frequency list and reverse vocabulary for vernacular forms. Lorna Stevenson and Jocelyn Wogan-Browne are continuing work on a companion concordance to the 'Katherine' Group (with appendices on the Wooing Group). The project was originally designed to provide two hard-copy concordances for ready reference in linguistic and literary study of these important early thirteenth-century texts. The possibilities of CD Rom publication for the whole data base (with both the completed and edited 'A' forms as well as the 'B' forms for which editing is in progress) are currently being investigated.

Jocelyn Wogan-Browne: AE18@liv.ac.uk
Lorna Stevenson:
c/o Canterbury Tales Project,
Centre for Sociolegal Studies, Linton Rd, Oxford

SOFTWARE DEVELOPMENTS

15. 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 system completed in Amsterdam is being used to produce tagged text. 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 1996. The corpus is expected to be in distribution within four years.

ICAME Journal 19: 151, 156-157)
Susan Pintzuk: sp20@york.ac.uk

16. Developing the Constraint Grammar Parser of English for the Analysis of Early English Texts

ENGCG, originally designed to deal with Present-day English, can relatively easily be developed so that it can be used to analyse Early Modern English texts automatically. Using early English texts, the ENGCG lexicon is updated and a specialised grammar written to make the morphological disambiguation grammar effective. Some manual editing is required prior to and after the analysis.

For further information, see 'Applying the Constraint Grammar Parser of English to the Helsinki Corpus', *ICAME Journal* (1995) 19: 23–48; 'Backdating the English Constraint Grammar Parser for the Analysis of English Historical Texts', a paper given at the XII ICHL Conference, Manchester, 13–18 August, 1995 (forthcoming).

Merja Kytö: merja.kyto@engelska.uu.se

Atro Voutilainen: avoutila@ling.helsinki.fi