Early English in the Computer Age

Explorations through the Helsinki Corpus

Edited by

Matti Rissanen
Merja Kytö
Minna Palander-Collin

Mouton de Gruyter
Berlin · New York 1993

Topics in English Linguistics
11

Editors

Jan Svartvik
Herman Wekker

Mouton de Gruyter
Berlin · New York

Early English in the Computer Age

Explorations through the Helsinki Corpus

Edited by

Matti Rissanen
Merja Kytö
Minna Palander-Collin

Mouton de Gruyter
Berlin · New York 1993

Mouton de Gruyter (formerly Mouton, The Hague)
is a Division of Walter de Gruyter & Co., Berlin.

♾ Printed on acid-free paper which falls within the guidelines
of the ANSI to ensure permanence and durability.

Library of Congress Cataloging-in-Publication Data

Early English in the computer age : explorations through the Helsinki corpus / edited by Matti Rissanen, Merja Kytö, Minna Palander-Collin.
p. cm. – (Topics in English linguistics ; 11)
Includes bibliographical references (p.).
ISBN 3-11-013739-9 (cloth : alk. paper)
1. English language – Early modern, 1500–1700 – Research – Data processing. 2. English language – Middle English, 1100–1500 – Research – Data processing. 3. English language – Old English, ca. 450–1100 – Research – Data processing. 4. English language – Discourse analysis – Data processing. 5. English language – Data bases – Finland – Helsinki. 6. Computational linguistics. I. Rissanen, Matti. II. Kytö, Merja. III. Palander-Collin, Minna, 1967– IV. Series.
PE1074.5.E37 1993
427'.00285 – dc20 93-27036
CIP

Die Deutsche Bibliothek – Cataloging-in-Publication Data

Early English in the computer age : explorations through the Helsinki corpus / ed. by Matti Rissanen ... – Berlin ; New York : Mouton de Gruyter, 1993
(Topics in English linguistics ; 11)
ISBN 3-11-013739-9
NE: Rissanen, Matti [Hrsg.]; GT

Typesetting and Printing: Arthur Collignon GmbH, Berlin.
Binding: Lüderitz & Bauer, Berlin.
Printed in Germany.

Preface

In the autumn of 1983 some members of staff and a number of post-graduate students of the English Department of the University of Helsinki got together to discuss the possibility of initiating a new departmental project. This project was intended to bring together the corpus research interests of the department and a number of individual pre- and post-doctoral studies. It was also considered sensible, for various reasons, that the project should make use of the new but rapidly developing computer applications in the field of the humanities.

As a result of this discussion, we began in 1984 to compile a large two-part computerized corpus, which was given the name *The Helsinki Corpus of English Texts: Diachronic and Dialectal.* To connect the study of the history of English with the research of present-day British dialects was well in accordance with our idea of approaching the change and development of language through synchronic variation. The diachronic part was completed in August 1991 and became available for international scholarly use in October the same year.

At present, the Helsinki Corpus is the only structured corpus of English stretching over the Old, Middle and Early Modern English periods. It consists of samples of English writing, from the eighth to the eighteenth century, mainly prose but also verse. Its total size is c. 1.5 million words. Each text sample is provided with parameter codings giving information on the text and identifying its author if known. In addition to the basic corpus, there are two supplementary corpora in preparation, one on Older Scots and the other on early American English. The corpus is particularly intended for diachronic studies of morphosyntax and lexis; with certain reservations, it can also help in historical studies of phonology, style, text and discourse. We hope to provide at least parts of the corpus with grammatical tagging in the future.

In recent years, problems of the size and representativeness of text corpora have been the topic of lively discussion. In compiling a diachronic corpus, these problems are multiplied. Our corpus, of course, does not represent all types and levels of written English of past centuries, although we have done our best to make our collection of texts as varied and

wideranging as possible. Furthermore, for many research purposes, the size of our corpus is too limited, particularly if the aim is to trace the development of a form or expression through successive synchronic stages of the language, paying attention to linguistic and extralinguistic factors affecting the distribution of variant forms. We acknowledge the restrictions of our corpus humbly but not too apologetically. We see our work as one of the first steps in the field of diachronic corpus compilation, which may encourage the creation of a series of new historical corpora.

At the moment promising projects are being launched in various parts of the world, aiming particularly at the preparation of Middle English and Late Modern English corpora. These projects will no doubt greatly benefit from the recent advances in theory and know-how in corpus linguistics and from the revolutionary development of software and hardware. We hope that the Helsinki Corpus can serve as a kind of macrodiachronic corpus and reference point for new corpora, which will be chronologically more focused and, consequently, will be more representative of the synchronic stages of the past. We are, of course, happy to share our knowledge and experience with the compilers of these corpora.

More than a dozen scholars from the English Department participated in the compilation of the diachronic part of the Helsinki Corpus, giving generously of their time. A number of advanced undergraduate students efficiently keyed in most of the material in circumstances which, particularly in the first years of the project, were far from ideal. (A more detailed statement of acknowledgements and the special contributions of the project group members is included in the Preface to the *Manual to the diachronic part of the Helsinki Corpus of English Texts* (compiled by M. Kytö, Helsinki University Press, 1991, 2nd ed. 1993).) The hero of the corpus project was, without the slightest doubt, Dr. Merja Kytö, who supervised the keying-in and proofreading of the texts, coordinated the team work, devised the database arrangements, guided visitors and was, in general, in charge of the everyday routine of the project.

The purpose of the present volume is to introduce the diachronic part of the Helsinki Corpus in a less technical way than has been done in the *Manual*, and to give examples of the varying possibilities offered by the corpus for the study of the history of English. The team responsible for the writing and editing of this book consists of the core group of the corpus project. Each member has participated in writing the introductory chapters for the various sections of the corpus: Leena Kahlas-Tarkka, Matti Kilpiö and Aune Österman for Old English; Saara Nevanlinna,

Päivi Pahta, Kirsti Peitsara and Irma Taavitsainen for Middle English; and Terttu Nevalainen and Helena Raumolin-Brunberg for Early Modern British English. Anneli Meurman-Solin and Merja Kytö have contributed the introductions for the supplementary Older Scots and early American English corpora, and Merja Kytö and Matti Rissanen are responsible for the General Introduction. In addition, this book contains eight pilot studies. The main purpose of these studies has been to illustrate the various ways in which evidence can be retrieved from the Helsinki Corpus; no far-reaching or conclusive results have been sought.

The present volume is the first product of a new research project launched at the English Department in 1990. The ultimate aim of this project is to produce another corpus-based volume which will contain a discussion of certain central issues in the development of English. Simultaneously, the new project will test the present version of the corpus and pave the way for a new, improved and extended version. It is also our purpose to find out to what extent recent theories of language change can be confirmed or questioned by corpus-based study.

We wish to thank Mr Mark Shackleton for revising the language of the introductions and articles in this volume and Miss Kirsi Heikkonen for expert help in the final stages of editing the volume. We are very much indebted to the Academy of Finland for a three-year research grant for the new project, to the University of Helsinki for additional funding and providing premises for the project, and to Mouton de Gruyter for publishing this volume.

Helsinki, July 1993 *Matti Rissanen*

Contents

General introduction 1
Merja Kytö and *Matti Rissanen*

Period introductions

Old English 21
Leena Kahlas-Tarkka, Matti Kilpiö and *Aune Österman*

Middle English 33
Saara Nevanlinna, Päivi Pahta, Kirsti Peitsara and *Irma Taavitsainen*

Early Modern British English 53
Terttu Nevalainen and *Helena Raumolin-Brunberg*

Older Scots 75
Anneli Meurman-Solin

Early American English 83
Merja Kytö

Pilot studies

Introduction 95
Matti Rissanen

Syntactic and semantic properties of the present indicative forms of the verb *to be* in Old English 97
Matti Kilpiö

Þær compounds from Old to Early Middle English 117
Aune Österman

The structure of Middle English similes of equality 139
Saara Nevanlinna

Genre/subgenre styles in Late Middle English? 171
Irma Taavitsainen

Toward the Modern English dichotomy between *every* and *each* 201
Leena Kahlas-Tarkka

On the development of the *by*-agent in English 219
Kirsti Peitsara

Periphrastic *do* in sixteenth- and seventeenth-century Scots 235
Anneli Meurman-Solin

"By and by enters [this] my artificiall foole ... who, when Jack beheld, sodainely he flew at him": Searching for syntactic constructions in the Helsinki Corpus 253
Merja Kytö and *Matti Rissanen*

Bibliography 267
Index of authors 291
Index of subjects 294

General introduction

Merja Kytö and *Matti Rissanen*

1. Purpose of the corpus

The main aim of the diachronic part of The Helsinki Corpus of English Texts (henceforth the Helsinki Corpus) is to support the variationist approach to the history of English, i.e., research based on extensive evidence provided by as many types, modes and levels of linguistic expression as possible. This kind of approach is of course traditional but it is also the most natural in historical studies of language, which cannot rely on native speaker informants or the student's own intuitions. We particularly hope that the Helsinki Corpus will encourage sociohistorical studies, which have offered most promising results in the last two decades or so. At the same time our corpus may help to evaluate and further develop recent theories of change, which are built upon both sociolinguistic considerations and other, more general aspects of change through variation.

Our corpus will probably be most useful for the studies of morphology, syntax and lexis. An attempt has been made to take the samples from good editions and to reproduce their spelling as accurately as possible (manuscript variants have not been included). The parameter codings appended to each sample, giving information on the type of the text and on the background of its author, may be helpful for students of historical stylistics. Finally, samples of continuous text varying from 2,500 to some 20,000 words may offer an adequate basis for the historical study of text and discourse.

At present, the usefulness of our corpus is diminished by the absence of grammatical tagging. This means that all searches must be based on words, or their parts or combinations. Programs suitable for tagging historical text material are being developed in various parts of the world, and we hope to start applying these programs to some subsections of our corpus in the near future. It is obvious, however, that equipping the entire corpus with even a fairly simple grammatical tag system would, with our present resources, be a matter of years of hard work.

2. Size and diachronic structure

The diachronic part of the Helsinki Corpus is divided into two entities, the basic corpus and the supplementary corpora. The basic corpus is systematically compiled and coded; it is the part intended for general scholarly distribution (for coding conventions and distribution formats, see the *Manual* [Kytö 1991]). Supplementary corpora have been, and are being, collected by individual scholars for their own research purposes; they follow the same typographical conventions and coding systems as the basic corpus and can be later appended to it for more general scholarly use. This arrangement should contribute to the continuous development and improvement of the corpus: the most important supplementary corpora under preparation, Older Scots by Anneli Meurman-Solin and early American English by Merja Kytö, will mean a decisive extension to the coverage of geographical variety of the Helsinki Corpus (see the introductions below).

The size of the basic corpus is c. 1.5 million words. Its organization follows the traditional division into Old English, Middle English and Early Modern (Southern) British English sections (see Table 1). Old and Early Middle English texts are grouped into century-long subsections (from mid-century to mid-century); later Middle English and Early Modern English are divided into subperiods of 70 or 80 years.[1] The rationale for these divisions is discussed in the period introductions below.

The first problem to be decided upon in compiling a corpus is its size. The target originally set for the Helsinki Corpus was one million words, on the model of the Brown and Lancaster-Oslo/Bergen corpora. It was soon apparent, however, that the corpus could not be compressed within the million-word framework, and we decided to let it grow naturally without rigid preset limits. This policy was in accordance with our general attitude to let practical and heuristic factors prevail over the demands of logic and symmetry in the shaping and structuring of the corpus. This kind of liberalism can be criticized, but at present it seems to be the most feasible method to follow in compiling a historical corpus. The reasons for this are discussed below, in this introduction and in the period introductions.

Admittedly, even 1.5 million words make too small a corpus for the purposes of many diachronic studies with past synchrony as their starting point. If the millennium is sliced into periods of a century, the average size of each subcorpus, by even division, is c. 150,000 words. At the

Table 1. The diachronic part of the Helsinki Corpus: Size and period divisions

Subperiod	Dates	Words
Old English		
OE1	– 850	2,190
OE2	850– 950	92,050
OE3	950–1050	251,630
OE4	1050–1150	67,380
		413,250
Middle English		
ME1	1150–1250	113,010
ME2	1250–1350	97,480
ME3	1350–1420	184,230
ME4	1420–1500	213,850
		608,570
Early Modern English, British		
EModE1	1500–1570	190,160
EModE2	1570–1640	189,800
EModE3	1640–1710	171,040
		551,000
Total (Basic corpus)		1,572,820
Supplementary parts in preparation		
Older Scots		c. 600,000
Early American English		c. 300,000

moment, the most extensive Old English subcorpus has c. 252,000 words, and the largest Middle English subcorpus c. 214,000 words. The Early Modern English subcorpora are more equal in size, between 171,000 and 190,000 words each.[2] These figures are substantial but the amount of linguistic evidence dwindles rapidly if the material is broken down according to various linguistic and extralinguistic parameters (e. g., the type of text, for which there are more than twenty possible values). It is our intention to expand and improve the corpus in the future by adding new samples and dropping less suitable ones – the time for a new issue might be, for instance, five years after the publication of the present version.

The results yielded by our corpus in its present form can be called diagnostic and should, of course, not be considered in any way definitive

in regard to the "reality" of the English language of the past. Our experience, however, is that the Helsinki Corpus is extensive enough to show fairly reliable and consistent trends of development in a large number of topics, and we hope to show this by the pilot studies included in this volume. The users can easily sharpen the picture given by our corpus with details obtainable from the texts themselves, from other corpora such as the Toronto Old English Corpus (cf. Healey – Venezky 1980), from printed concordances, and so forth.

Our avoidance of a rigid symmetrical structure to the corpus can also be seen in the choice and size of the individual samples of text. The basic procedure was first to decide what genres and texts were felt necessary to be included in each section of the corpus and then to see how long the samples could be in order to keep the section and the subsections reasonably comparable in size. Short texts were included *in toto*; the extracts sampled from longer texts vary considerably in size. Certain prominent texts and authors (*Beowulf*, Alfred, Ælfric, Chaucer, Rolle) in the history of English have been extensively sampled. In the Early Modern English section, with a wider variety of text types to include, the individual extracts tend to be shorter than in Old and Middle English. In general, we have aimed at samples of c. 10,000 words in the early subperiods and between 2,500 and 5,000 words in late Middle English and Early Modern English. At least two separate extracts are normally included in each text sample.

Although it is important for the user of the corpus to be aware of the varying length of the samples, with the calculation programs available this variation does not present major problems for quantitative analysis. Furthermore, it is relatively easy for the user to exclude samples or cut them to size according to the demands of the research topic, as the text files are easy to edit both in the mainframe and diskette versions.

3. Coding the texts

The Helsinki Corpus differs from previous multipurpose corpora known to us in that each text sample is described by a number of parameters comprising a set of two or more values. As can be seen in the following sample taken from our corpus, the parameter values give shorthand information on the text and its author. The codes are introduced in COCOA format (cf. the Oxford Concordance Program).


```
<Q E2 XX CORP HARLEY>
<N LET TO HUSBAND>
<A HARLEY BRILLIANA>
<C E2>
<O 1570–1640>
<M X>
<K X>
<D ENGLISH>
<V PROSE>
<T LET PRIV>
<G X>
<F X>
<W WRITTEN>
<X FEMALE>
<Y 20–40>
<H HIGH>
<U X>
<E INT UP>
<J INTERACTIVE>
<I INFORMAL>
<Z X>

<P 1>
[}TO MY DEARE HUSBAND S=R= ROBART HARLEY, KNIGHT OF THE BATHE.}]                #

  S=r= - Docter Barker has put my sister into a cours of                        #
ientell fisek,
which I hope by God's bllsing will doo her much good. My sister
giues you thankes for seending him to her. I pray you remember
that I recken the days you are away; and I hope you are nowe
well at Heariford, wheare it may be, this letter will put you                   #
in minde
of me, and let you knowe, all your frinds heare are well; and                   #
all
the nwes I can seend you is, that my Lo. Brooke is nowe at                      #
Beaethams
Court. My hope is to see you heare this day senet, or
to-morrowe senet, and I pray God giue vs a happy meeting, and
presarfe you safe; which will be the great comfort of

Your most true affectionat wife, Brilliana Harley
```

This is Lady Brilliana Harley's letter to her husband written in 1625, with parameter codings added to the beginning of the text.[3] The top line in the code column can be used for the reference to each example taken from the text with WordCruncher, the Oxford Concordance Program, or other concordance programs.[4] The code values underneath indicate, among other things, that the text is a private letter (T) representing an

interactive type of discourse (J) in an informal setting (I). The author is female (X), between 20 and 40 years of age (Y), and has a high social position (H). The writer and addressee are on intimate terms but have unequal social standing (E). The parameter C defines the subperiod in which this text was written.

The purpose of our coding is to offer the user of the corpus basic information on the sample in question. Such retrieval programs as the Oxford Concordance Program (*OCP*, 1984, 1988) can use the coding system for searches directed to only those samples which fulfil a predefined set of parameter values. The parameter coding enables the users of the corpus to choose between two approaches. One approach is to collect all the occurrences of the structure or lexical item under scrutiny, in the entire corpus or a part of it, and then observe the distribution of the instances with regard to the various parameters (dialect, text type, author, and so forth). Alternatively, users can restrict the search to samples fulfilling a certain set of values (e.g., private letters written by young women between 1500 and 1640; religious treatises representing the East Midland dialect in 1250–1500) and contrast the results with distributions found in samples differently defined. In actual practice, users probably apply both methods side by side: they most likely begin with an overall survey of the occurrences of the form or construction under scrutiny, and after preliminary findings may continue with a check of the influence of varying combinations of constraints.

An attempt to define the parameter values for each text sample has been both rewarding and frustrating work. The further back in time we go, the hazier the contours of the available text material become; for one thing, the anonymity of medieval texts is the rule with relatively few exceptions. We have done much philological detective work, particularly in the case of our Old and Early Middle English samples. In many cases, however, we have been compelled to resort to the X value, which indicates either "not applicable or irrelevant to this sample"[5] or, lamentably often, "not known" or "information too uncertain or inaccurate to be coded". We are also certain that with the increase of philological and linguistic knowledge we will have to revise many of the parameter values assigned to the texts.[6] Nevertheless, the pilot studies completed so far have proved our coding to be a satisfactorily powerful tool in the variationist analysis of corpus texts, and we sincerely hope that all compilers of corpora in the future would consider providing the text samples with textual parameter coding.

4. Selection of texts

In compiling a diachronic multipurpose corpus, the following four aspects should be taken into consideration. The last three of these criteria are of course applicable to all corpora.

- Chronological coverage: the corpus should be representative of all parts of the period(s) it is intended to cover.
- Regional coverage: the corpus compiler should pay attention to the regional varieties of the language.
- Sociolinguistic coverage: the texts of the corpus should be produced by male and female authors representing different age groups, social backgrounds and levels of education.
- Generic coverage: the corpus should contain samples representing a wide variety of genres or types of text.

In the following, we will give a brief survey of the general principles followed in observing these criteria in our selection of texts: more detailed accounts are given in the period introductions below. It may be added that our texts are derived from editions (for purely practical reasons) and that we avoid translated texts in Early Modern English (with a few well-defined exceptions) and concentrate mainly on prose texts. We have given high priority to a diverse selection of different types of texts, even at the risk of upsetting the regional dialect balance of the samples in our earlier subperiods.

4.1. Chronological coverage

An even chronological spread of texts is of course of prime importance to the compiler of a diachronic corpus: the very idea of this kind of corpus is to give the student an opportunity to map and compare variant paradigms (variant fields) in successive synchronic stages in the past.

In addition to the problem of lack of texts in the earliest stages of the language, the compiler of a diachronic corpus also has to face the problem of how to shape the chronological ladder of the successive synchronic stages. With the Helsinki Corpus we have, even here, let practical factors outweigh systematic and symmetrical solutions. As mentioned above, and as shown by Table 1, our Old and Early Middle English subperiods are a century long, our Late Middle English and Early Modern English ones, 70 to 80 years.

In future versions of the corpus we may consider reducing the length of our subperiods. It must be kept in mind, however, that in Old and Early Middle English an accurate dating of the texts is impossible, and the chronological definition of the stage of language represented by the early samples is further complicated by the long and uncertain manuscript history of many texts. The difference between the date of the manuscript included in the corpus and the (often hypothetical) date of the original may well be over a hundred years and cross a subperiod division. We have tried to cope with this problem by using, when necessary, two code values for the definition of the subperiod of the sample. OE2/3, for instance, indicates that the original text probably dates from OE2 (850–950), while the manuscript included in the corpus was written in OE3 (950–1050). Fairly often, the code value denoting the subperiod of the original version of the text has been indicated by an X, signalling our ignorance or uncertainty about the date of composition. In our rough grouping of texts into subsections, the date of the manuscript has been decisive.

Ideally, each subsection should contain the same amount of text. For obvious reasons, this symmetry has been impossible to attain. In the earliest Old English century, around the time of the Conquest, and in the thirteenth and early fourteenth century, the amount of English text material is so small that these subperiods (OE1, OE2, OE4; ME1 and ME2) are underrepresented. We have not been willing to compensate for this scantiness by overlong samples from the few extant texts as we feel that this would have been even more harmful to the comparability of the subsections of the corpus than the uneven chronological distribution.

4.2. Regional dialect

It is impossible to think of a diachronic corpus which would not pay attention to regional dialect distribution in the periods preceding the establishment of the standard. In the Helsinki Corpus, all samples up to 1500 have been given dialect or localization parameter values, although in the case of many fifteenth-century texts the definition, based on external evidence, simply signals that the sample represents some stage of development in the Southern standard. In defining the dialects of later Middle English samples, the information provided by the *Linguistic Atlas of Late Medieval English* (*LALME* 1986) has been indispensable. The dialect codings of most earlier Middle English texts are based on the definitions of the *Middle English Dictionary* (*MED* 1954–).

One problem in defining the dialects, as well as the dates of composition, of the early samples is that these definitions must be based on earlier linguistic research. In most cases, too little is known about the authors and the provenance and manuscript history of the texts to rely on extralinguistic criteria. Thus, a certain degree of circularity is unavoidable in dialect studies based on our parameter values. We hope, however, that the easy access to the earliest text material provided by the corpus will help scholars to revise and sharpen their theories and assumptions on Old and Early Middle English dialects.[7]

In Early Modern English, all texts selected represent the Southern standard; collecting a dialect corpus from this period would need a completely new project and much additional research. The Scots and American English supplementary corpora will, however, make it possible to observe and compare distributions between major regional varieties even in sixteenth- and seventeenth-century texts.

4.3. Sociolinguistic factors

While the regional dialect parameter is important only in Old and Middle English, the importance of the parameters giving sociolinguistic information on the authors is greater for the later sections of the corpus. These parameters apply from Middle English on: even if we possess some information on such Old English authors as King Alfred or Archbishop Wulfstan, this information is too occasional to offer a basis for sociohistorical generalizations.

Particularly in the late Middle and Early Modern English sections we have done a great deal of work to give reliable values to the parameters defining the age (in twenty-year age groups) and social status (on a simple high/professional/other scale) of the authors. In describing correspondence, private and official, we have also paid attention to the relationship between the sender and the receiver, coding it as "intimate" or "distant". The writer and the addressee may be ranked as "equal", or the letter may be addressed to a person in a higher ("up") or lower ("down") social position. All official letters are "distant" by definition and those by core family members, "intimate". A wife's letter to her husband, children's letters to parents and all letters to the king or queen by his/her subjects are coded as "up", two members of the gentry are "equal", and so on.

In our collection of scientific and instructive writings, we have included (and coded) treatises intended for either "professional" or "non-professional" readers, as this distinction may be of considerable import for the

quality of the vocabulary and the general level of expression. Finally, we have observed the difficult and elusive concept of formality in selecting and describing our samples. Formality has been defined mainly on the basis of the discourse situation. Thus, sermons, trial records and official correspondence have been coded as "formal" and private correspondence and comedy as "informal". The latter label has also been applied to Early Modern English light fiction, which is intended for the entertainment of the reader.

4.4. Types of text

Text type categorization is a highly relevant but also the most difficult structural aspect in the compilation of a corpus. It seems that no theoretically satisfactory classification by text type has so far been developed for this purpose; much more research is needed in this field. A glance at our text type list (Table 2) clearly shows that we have even here followed heuristic rather than logical principles (for distinctions between text types, such as handbooks versus science, see the Period introductions).

One advantage of diachronic text type definitions, in comparison to chronological and dialectal ones, is that they can be primarily based on extralinguistic criteria, and the risk of circularity is in this way diminished. These criteria mainly pertain to the subject matter and purpose of the text, to the discourse situation and to the relationship existing between the writer and the receiver. Earlier studies on register, formality, discourse types, etc., have helped us considerably in our general approach and definitions.[8]

It should be obvious from this account that our text type codings do not indicate linguistic or discursive uniformity of the texts grouped under one and the same category. Therefore, the user of the corpus should keep an open mind about grouping the samples on textual principles which may differ from our codings. It is fair to assume, however, that the texts included in the same category show some common features. Furthermore, some classification is necessary to offer a basis for the study and further discussion of the typological features of the English texts of the past. And finally, text type coding is an efficient way of indicating to users what kinds of written English they can expect to find in the corpus.

It has been our aim to give as rich and many-sided a selection as possible of the writings produced in all subperiods covered by our corpus. Our samples contain both literary and nonliterary texts, with an obvious bias towards the nonliterary genres, texts of the public and private

Table. 2. Text type

Old English	Middle English	Early Modern English
Law	Law	Law
Document	Document	
Handbk. Astronomy	Handbk. Astronomy	
Handbk. Medicine	Handbk. Medicine	
	Handbk. Other	Handbk. Other
Science Astronomy		
	Science Medicine	Science Medicine
		Science Other
		Educat. treatise
Philosophy	Philosophy	Philosophy
Homily	Homily	
	Sermon	Sermon
Rule	Rule	
Relig. treatise	Relig. treatise	
Preface/Epilogue	Preface/Epilogue	
	Proceed. Deposition	
		Proceed. Trial
History	History	History
Geography		
Travelogue	Travelogue	Travelogue
		Private diary
Biog. Saint's Life	Biog. Saint's Life	
		Biog. Auto
		Biog. Other
Fiction	Fiction	Fiction
	Romance	
	Drama Mystery	
		Drama Comedy
	Private correspondence	Private correspondence
	Non-private correspondence	Non-private correspondence
Bible	Bible	Bible
X	X	

Abbreviations

Handbk. = handbook
Educat. = educational
Relig. = religious
Proceed. = proceeding
Biog. = biography
Biog. Auto = autobiography
Drama Mystery = mystery play

domain, and so forth. The majority of the texts are in prose, to avoid the effect of verse form on the constructions; verse texts have been included only when they have been of particular importance (Old English poetry, *Layamon's Brut*, *The Ormulum*), or when the only existing samples of a text type in a (sub)period are in verse (earlier Middle English romances, Late Middle English and early sixteenth-century drama, and so on).

The major frustration in variationist studies of the early periods is the total lack of first-hand textual evidence on spoken language. We have tried to overcome this problem by including in our corpus a number of texts which can be used for obtaining information on the forms and constructions typical of speech. Some of the special characteristics of spoken expression are reflected, for instance, by Late Middle English and Early Modern English drama and private letters (many of these letters reflect the oral mode of expression), and by Early Modern English sermons, trial records and even the dialogue in fictitious anecdotes and jests. By studying the distribution patterns in these texts, and by juxtaposing the frequencies with those yielded by "non-speech-based" texts, we can obtain information on the ways of expression which were either favoured or avoided at the spoken level of language.

We have found the concepts of the oral and literate modes of expression, as expounded by, e. g., Traugott and Romaine (1985), useful in defining the relative distances of texts from spoken language. All the text types mentioned above do not necessarily represent the oral mode of expression, although they may stand in a relatively close relationship to spoken language. The sermon with its narrow scope of topic and a bias towards the formulaic and the ritual may be further removed from natural spoken expression than, say, a humorous narration or an intimate private letter, or even a formal presentation by an uneducated person possessing a very narrow repertoire of "styles". Similarly, the highly formal and "unnatural" discourse situation of a court trial must be borne in mind when conclusions are drawn from the dialogues in trial records.

A textual feature which seems highly relevant in our search for information on the spoken level of language is the possible interactivity of the discourse. Typically, interactive texts are plays, court trials, discussions and debates. We have also, for better or worse, treated correspondence as interactive. Most letters are worded with the expectation of a response, and particularly in private correspondence the attitude of the author is more conversational and first/second person oriented than in most other non-speech-based text types.

All in all, there are 22 text types, with further subcategories, represented in the corpus. Of these types, nine occur in all the three main sections of the corpus, though not necessarily in each of the subsections. These are law, handbooks, science, philosophy, history, travelogue, biography, fiction, and the Bible.

It is only to be expected that not all text types are represented in all subperiods in the Helsinki Corpus. Furthermore, the amount of text representing a certain genre may be too small to support conclusions, particularly if low-frequency items are studied. In the case of some text types (e. g., documents), the coverage of the corpus can be extended later; some other genres will necessarily remain more or less nondiachronic. To diminish the disadvantages of this "generic noncontinuity", and to offer the users of the corpus a novel approach to textual material, we have, experimentally, grouped the types of text into the following larger entities called diachronic text prototypes. (The asterisk indicates that not all representatives of the type of text in question belong to one and the same prototypical category.)

statutory (STA):
law, document*
secular instruction (IS):
handbook, science (astronomy*, medicine*), philosophy*, educational treatise*
religious instruction (IR):
religious treatise*, homily, rule, preface*, sermon
expository (EX):
science (astronomy*, medicine*, other), educational treatise*
nonimaginative narration (NN):
history, biography (saint's life, autobiography, other), religious treatise*, Middle English secular lyric*, travelogue*, diary
imaginative narration (NI):
fiction, romance, travelogue*, geography

At least a few samples representing each of these six categories can be found in Old, Middle and Early Modern English. The purpose of even this categorization is primarily practical; it follows, to some extent, the discussion in Werlich (1983). The distinction between secular and religious instruction is useful in view of the Old and Middle English samples; "imaginativeness" separates fiction, romances, etc., from, e. g., histories and biographies. Two important categories, argumentation and description, are missing; this is because it would have been extremely difficult

to find texts in Old and Early Middle English in which these categories did not overlap with the others mentioned above.[9]

The diachronically prototypical categories necessarily contain heterogeneous material and so far no detailed studies are available to test their usefulness. It is to be expected that our groupings and codings will need considerable reorganization in the future, with the increase in our knowledge of the diachronic aspects of the generic features of texts. It would seem, however, that this kind of classification might help, for instance, in the study of the linguistic characteristics typical of narration or instruction, in the special lexical features of religious language, and so forth.

The following text types, no less important, remain outside the diachronic prototypes: Old English verse and some Middle English secular lyrics; some documents; some philosophical treatises; some prefaces and epilogues; the Bible; depositions; drama (mystery and comedy plays); correspondence; and trials.

Diachronic continuity is also aimed at in our decision to include in our corpus several translations dating from different centuries of two texts. These are, not surprisingly, the Bible and Boethius' *De Consolatione Philosophiae*. Our Bible extracts come from six Old English, three Middle English and three Early Modern English versions. Boethius is sampled from Alfred's and Chaucer's translations and from three Early Modern English renderings. Otherwise, we have not aimed at a systematic collection of translated texts, although we have a special code for "translation" and another indicating the source language.

As can be seen in the above discussion we have not included "style" as a criterion for selection or as a parameter to be coded. We have made this difficult but necessary decision because diachronic stylistics has been as yet too little studied to enable an adequate classification. Furthermore, we feel that it is impossible to define the style of a text on extralinguistic criteria, and using linguistic criteria would inevitably lead to circularity. It is our wish that our corpus, with its textual and sociolinguistic parameter codings, be sufficiently representative of stylistic varieties to offer a basis for the study of the linguistic features of style.

5. Distribution versions and coding conventions

The diachronic part of the Helsinki Corpus is available in three main versions: 1) the 242 text files containing one text (or group of related texts) per file; 2) a one file version integrating all individual files (the order of files follows the chronology and the order of texts in the text type chart); 3) a WordCruncher version in three formats (one WordCruncher book containing the entire corpus; three books, each containing one main period; and eleven books, each containing one subperiod). In addition, our corpus is included in the CD-ROM disk "ICAME Collection of English Language Corpora".

When keying in the text, we followed the original editions as closely as possible. Typographical conventions (italics, superscript, accents), headings, foreign words or passages, runes, emendations, and editors' and our own comments have been systematically coded (for details, cf. Kytö 1991 [2nd ed. 1993]).

Notes

1. The following terminology is applied to the parts of the Helsinki Corpus: diachronic part, dialectal part; Old English, Early and Late Middle English, and Early Modern English periods/sections; subperiods/subsections, 850–950, 1350–1420, 1570–1640, and so on. The subperiods are referred to as OE1, ME2, EModE3, etc., or, for short, O1, M2, E3, and so forth.
   The word counts given in the General Introduction and the Period Introductions (Old, Middle and Early Modern British English) exclude the portions of text coded as instances of "foreign language", "editor's comment" or "our comment".
2. The sub-period 1640–1710 contains less text than the other two Early Modern English ones because it does not contain samples from the Bible.
3. The typographical conventions are explained in Kytö (1991).
4. So far our pilot studies are based on the use of WordCruncher and, to a lesser extent, on the Oxford Concordance Program.
5. As, for instance, the M and K parameters in Lady Brilliana Harley's letter, above, which specify the date of the extant manuscript and the contemporaneity of this manuscript with the original text, or the G and F parameters which give information on texts translated from other languages.
6. It is evident that work of these dimensions always contains printing errors and mistakes. We are grateful for all comments and lists of errors sent to us (address: Professor Matti Rissanen, Department of English, P.O. Box 4, Hallituskatu 11, FIN-00014 University of Helsinki, Finland). The following errors in parameter codings have been noticed so far (April 1992): *Dame Sirith*: <D SL> read <D SO>; *Interlude* (*Appendix to Dame Sirith*): <D SL> read <D EMO>; *Kyng Alisaunder*: <Q M2 NI ROM KALEX> read <Q M2/3 NI ROM KALEX>, <C M2> read <C M2/3>, <M 1250–1350> read <M 1350–1420>, <K CONTEMP> read <K NON-CONTEMP>. The present volume was completed, before these errors were detected.
7. The *Linguistic Atlas of Early Middle English*, in preparation, will of course decisively increase our knowledge of the earlier Middle English dialects. The Toronto Old English Corpus has very much improved the possibilities for extensive studies of Old English dialects.

8. Among many important studies in the field, we would particularly like to mention Romaine (1982), Werlich (1983), Halliday–Hasan (1985), Traugott–Romaine (1985), Milroy–Milroy (1985), and Biber (1988).
9. This is, to a certain extent, due to the internal variation within the samples. Implementing systematic intratextual parameter coding has been beyond our resources.

References

Biber, Douglas
1988 *Variation across speech and writing*. Cambridge–New York: Cambridge University Press.

Halliday, M. A. K. – Ruqaiya Hasan
1985 *Language, context, and text: Aspects of language in a social-semiotic perspective*. Geelong: Deakin University Press.

Healey, Antonette diPaolo – Richard L. Venezky
1980 *A microfiche concordance to Old English*. (Publications of the Dictionary of Old English 1.) Toronto: The Pontifical Institute of Mediaeval Studies.

Kytö, Merja (comp.)
1991 *Manual to the diachronic part of the Helsinki Corpus of English Texts: Coding conventions and lists of source texts*. (2nd ed. 1993.) Helsinki: Department of English, University of Helsinki.

LALME=McIntosh, Angus – M. L. Samuels – Michael Benskin – Margaret Laing – Keith Williamson
1986 *A linguistic atlas of late mediaeval English*. Aberdeen: Aberdeen University Press.

MED=Kurath, Hans – Sherman M. Kuhn et al. (eds.)
1954– *Middle English dictionary*. Ann Arbor, Michigan: University of Michigan Press – London: Geoffrey Cumberlege and Oxford University Press.

Milroy, James – Lesley Milroy
1985 "Linguistic change, social network and speaker innovation", *Journal of Linguistics* 21: 339–384.

OCP=Hockey, Susan – Ian Marriott, (comps.)
1984 *Oxford Concordance Program. Users' manual, version 1.0*. Oxford: Oxford University Computing Service.

OCP=Hockey, Susan – Jeremy Martin, (comps.)
1988 *Oxford Concordance Program. Users' manual, version 2*. Oxford: Oxford University Computing Service.

Romaine, Suzanne
1982 *Socio-historical linguistics, its status and methodology*. (Cambridge Studies in Linguistics 34.) Cambridge–London: Cambridge University Press.

Traugott, Elizabeth Closs – Suzanne Romaine
1985 "Some questions for the definition of "style" in socio-historical linguistics", *Folia Linguistica Historica* VI: 7–39.

Werlich, Egon
1976 *A text grammar of English*. (Uni-Taschenbücher 597.) Heidelberg:
1983 Quelle–Meyer.

WordCruncher
1987 *WordCruncher. Text indexing and retrieval software, version 4.1*. Provo, Utah: Brigham Young University – Electronic Text Corporation.
1989 *WordCruncher. Text indexing and retrieval software, version 4.3*. Provo, Utah: Brigham Young University – Electronic Text Corporation.

Period introductions

Old English

Leena Kahlas-Tarkka, Matti Kilpiö and *Aune Österman*

1. Introduction

The texts representing the Old English period are divided into four subsections. Each of these covers a century with the exception of the earliest (OE1), which consists of writings prior to the year 850. Both prose and verse are included, although verse texts are in a clear minority. The size of the subsections, as well as the amount of prose and verse, are indicated in Table 1.

Table 1. The Old English period in the Helsinki Corpus: A quantitative overview

Subperiod	Words	%	Prose		Verse	
OE1 – 850	2,190	(0.5)	1,960		230	
OE2 850 – 950	92,050	(22.3)	91,680		370	
OE3 950 – 1050	251,630	(60.9)	174,010		77,620	
OE4 1050 – 1150	67,380	(16.3)	67,380		–	
	413,250	(100)	335,030	(81.1)	78,220	(18.9)

The average size of the text samples from longer texts is between 5,000 and 10,000 words. When one and the same text is represented in the corpus by different versions (e. g., the Bible, *Gregory's Dialogues*, or the *Chronicle*), the samples cover different passages of the text. This procedure has the obvious advantage of providing a larger sample of one particular text, but excludes the possibility of comparing different versions.

2. Chronological coverage

The starting point of the Helsinki Corpus being variation, the text material should be able to give as representative a picture of the Old English

period as possible and there should of course be enough material for comparison. As can be seen from Table 1, the texts are not evenly distributed between the subperiods: for obvious reasons, the subsections OE1 and OE4 are very small. Practically all extant pre-850 texts have been included: *Cædmon's Hymn, Bede's Death Song, Ruthwell Cross, The Leiden Riddle*, and the earliest documents (Birch 451, Harmer 1, 2, 3, 5, Robertson 3). For the post-Conquest OE4 subperiod relatively limited material is available: late annals of the *Chronicle*, the *Vision of Leofric*, and some documents (*William's Laws*, Robertson's *Appendix* 1, 3, 4). In addition, some texts which appear in late manuscripts, even though the originals are earlier, have been included in this subperiod. (See also below for an account of dating.)

On the basis of the number of words Old English poetry may seem overrepresented. Its importance for Old English studies cannot, however, be disputed, and therefore a relatively large selection of verse has been included to provide scholars with a representative corpus of poetic language.

3. Dating

Dating the Old English texts and grouping them according to the four-part division of the corpus causes a problem of its own. We have adopted a cautious and conservative approach, relying on previous scholarship and fully acknowledging the difficulties involved. Quite a few texts can be dated by historical evidence (*Bede's Death Song*, Wærferth's translation of *Gregory's Dialogues*, Alfredian translations, the works of Ælfric and Wulfstan). The entries of the *Chronicle* and the battle poems can be given at least *termini a quo* on the basis of the events they describe (Amos 1980: 1–2). But only a part of the corpus is datable on external or internal evidence.

The texts have been grouped into subsections according to the date of the manuscript sampled. In many cases, the date of the composition of the original version and that of the extant manuscript are different. If both are known, a double coding has been given (e. g., O2/4 for MS C of *Gregory's Dialogues*). More often, the date of composition cannot be established with any certainty. In those cases it is indicated with X (unknown), e. g., OX/2 for *Ine's Laws*. *The Blickling Homilies* included in the corpus have been coded O2/3, *Wulfstan's Homilies* either O3 or

O3/4, *Martyrology, Marvels* and *Alexander's Letter* O2/3, *Chad* O2/4, and *Gregory's Dialogues* MS H O2/3. As noted above, this principle of grouping the texts has also increased the amount of text for O4, which would otherwise have been considerably smaller.

A similarly cautious attitude towards dating Old English poetry has been adopted. Apart from the pre-850 poems, *The Battle of Brunanburh* (O2), and *The Meters of Boethius* (O2/3), all verse texts have been coded OX/3. Thus no stand has been taken regarding various attempts at dating some of the Old English poems with more precision.

4. Types of text

The selection of the Old English samples has been made with attention to a number of text typological features. As can be seen in Table 2 in the appendix (see also Table 2 in the General Introduction), the number of text types in the Old English corpus is smaller than in the Middle or Early Modern English corpora. In our period the written language had not yet realized its full potential in all walks of life. Society was developing, still creating its institutions. Gradually, however, learning started to spread through educational activities undertaken by the Church and through educational programmes launched by individual rulers, most notably by King Alfred. These processes are reflected in the increasing variety of texts towards the end of the Old English period. It must also be remembered that we have no idea of the amount of literature lost in the course of centuries.

Our selection of texts and their division into various types is relatively conventional and needs few comments. The law texts represent the category of statutory prose, while the documents are mostly wills and definitions of boundaries and are intended for individual or otherwise specific situations; thus they do not seem to contain the generalizing power required of the term "statutory". Medical recipes, of which there is no shortage in Old English, are typical representatives of the handbook and secular instruction categories. Astrological writings (prognostications) are also classified as handbooks, but *Byrhtferth's Manual* and *Ælfric's De Temporibus Anni* are regarded as scholarly treatises and thus labelled as "scientific". The *Manual* is clearly instructive, but *De Temporibus* can be regarded as a representative of expository writing.