MEG-C

Corpus Manual - version 2011.1

Merja Stenroos and Martti Mäkinen
February 2011

1. Preliminaries
    1.1. Introduction
    1.2. The purpose of this manual
    1.3. Notes on terminology
2. Principles of compilation
    2.1. The scope of the corpus
    2.2. The selection of texts
3. Processing a manuscript text from original/microfilm to a text file
    3.1. The selection of tranches
    3.2. Transcription routines
    3.3. Description of transcription conventions
        3.3.1. Spelling and abbreviations
        3.3.2. Punctuation
        3.3.3. Word and line division
        3.3.4. Foliation
        3.3.5. Coding for layout, corrections and commentary
    3.4. List of symbols used in the transcription
    3.5. Editorial decisions and interpretation in the Corpus
4. Different versions of the corpus
5. Searches and recommended software
    5.1. The MEG-C Concordance version
    5.2. Using the Concordance files with AntConc 3.2.1
6. Feedback
7. Updates
References

1. Preliminaries

1.1. Introduction

The *Middle English Grammar Corpus* (MEG-C) is a purpose-built text corpus consisting of samples of English texts from the period 1300-1500. All texts have been transcribed from the manuscript or from facsimile reproductions of the manuscript. The Corpus is intended both as an independent research tool and as the basis for a new description of linguistic variation and change in Middle English (a 'Middle English Grammar'). Shorter texts are included in their entirety and longer ones in 3,000-word samples.

The intention has been to include as many of the texts mapped in the *Linguistic Atlas of Late Mediaeval English* (McIntosh, Samuels and Benskin 1986, henceforth LALME) as the team is able to access; this will amount to 900-1,000 texts. In addition, the corpus will contain a large number of other texts from the same period. The first non-LALME groups of texts to be incorporated will consist of legal documents and Wycliffite writings. Version 2011.1 consists of 410 texts, all of which were mapped in LALME.

The corpus will form the main source material for further work within the Middle English Grammar Project. For that purpose, all the text samples are entered into a database together with information about extralinguistic variables such as date, genre, register and script; each word will be analysed into its spelling components and linked to headwords representing both Present-Day English and the immediate source language before Middle English (e.g. Old English or medieval French). This process is labour-intensive but will in the end, it is hoped, make possible a very thorough analysis of Middle English written variation.

In the meantime, the corpus is made available to the research community in its unannotated form. A preliminary version was made available on the project website in December 2007, and further versions were launched in April 2008 (MEG-C 1.0) and December 2009 (MEG-C 2009.1). The corpus is updated regularly as more texts and functionalities are added; however, each published version will remain available (see 7 below). Apart from the present Manual, the corpus is accompanied by a Catalogue of Sources, which is also updated for each version.

The corpus itself is produced in three versions designed for different uses. The **Base** version consists of .txt files with extensive coding and comments, and gives the fullest information about manuscript reality; it should be used as a reference version together with the others. The **Readable** version is produced as .html and .pdf files and is meant for reading rather than for electronic searching. The present version of MEG-C introduces the **Concordance version,** which consists of .txt files designed to be analysed using a concordancing programme or other corpus software. This version is tailored to suit a particular concordancing programme, AntConc 3.2.1., but it may be used with other programmes as well (see 5 below).

back to top

1.2. The purpose of this manual

This manual describes the sampling, transcription conventions and presentation of the corpus. It also gives brief instructions for the use of the concordancing programme AntConc for analysing the corpus files in the new _conc format (text files designed for use with a concordancer). The purpose is to provide users with the practical information needed for making use of the corpus. The manual does not discuss the wider research context nor the applications of the corpus within the Middle English Grammar Project; these questions will be addressed in the MEG Introduction.

back to top

1.3. Notes on terminology

The basic informant in the Middle English Grammar Project survey is the **scribal text**. The scribal text is defined in LALME (I: 8) as 'any consecutive written output that is a single text in the literary sense, or a part of such a text, and written by a single scribe'. The individual scribal texts, as defined in the MEG-C Catalogue, most often correspond to the texts underlying each **Linguistic Profile** in LALME. However, some of the LALME Linguistic Profiles are based on more than one strictly-defined scribal text (see 3.1. below). In such cases, the underlying text is split into individual scribal texts in MEG-C. Longer scribal texts are included as 3,000-word **samples**, which normally consist of two 1500-word **tranches**.

The word **text** may here be applied either to the scribal text or the sample. We try to avoid using the term in the sense of a literary text (e.g. *Piers Plowman* or the *Prose Brut);* here, the term **work** is preferred.

As this manual deals with original manuscripts, manuscript transcriptions and, occasionally, manuscript editions, it is important to distinguish between **scribes, editors** and **compilers.** A scribe is the person who committed the words of a manuscript to paper/vellum/parchment, i.e. the actual writer of the physical text as it survives to us. An editor is the person who has prepared a (printed) edition of a medieval manuscript text. By the term compiler we refer to ourselves, as in 'the compilers of this corpus'. The corresponding adjectives are, respectively, **scribal, editorial** and the somewhat unwieldy **compilatorial.**

back to top

2. Principles of compilation

2.1. The scope of the corpus

A corpus is sometimes distinguished from a text archive in that it is based on specific pre-defined principles of compilation, against which search results may be evaluated, rather than being simply a collection of texts that happen to be available.

The selection of texts for the present version of MEG-C is defined by a single external criterion: inclusion as a mapped text in LALME. As other texts are included in later versions

of the corpus, these will form distinguishable sub-corpora, marked with different code letters and subject to their own criteria of inclusion.

The finished corpus of LALME mapped texts will potentially consist of all the texts that are included in the Linguistic Profiles section of LALME, with the exception of two groups: 1) texts localized in Scotland and 2) early texts that fall outside the main chronological span of LALME and are also included in the *Linguistic Atlas of Early Middle English* (LAEME) at the University of Edinburgh.

The geographical scope of the Middle English Grammar Project is limited to England and Wales. This is not because the Scottish material would be uninteresting, but rather because it is felt to be a whole field of study of its own. For medieval Scots materials, the user is referred to the *Linguistic Atlas of Older Scots* (LAOS) at the University of Edinburgh.

The chronological scope of the main LALME material is ca 1325-1500. However, during the compilation of LALME, a separate survey of earlier materials was not yet envisaged. A small group of thirteenth-century texts was therefore also included, on the grounds that the dialectal material they provided was too important to ignore (LALME I: 3). These texts are not included in MEG-C, as a very full compilation of text materials from the earlier period (1150-1325) is now available in LAEME.

Apart from these two excluded groups, all texts listed in the LALME Linguistic Profiles section will, as far as possible, be included, whether or not they have been assigned a specific grid reference to the map. Thus, texts simply labelled as 'Northern' are included if they are represented by a Linguistic Profile in LALME. In practice, it is not envisaged that the corpus will ever contain every single text. Shelf marks and repositories have in some cases changed since the LALME survey, and some texts have become difficult or impossible to trace; other texts are difficult or impossible to access. The main principle of compilation is thus a relatively flexible one: the corpus will represent as large a proportion as possible of the texts localized in LALME, excluding the Scottish and Early Middle English texts.

back to top

2.2. The selection of texts

The 2011.1 version of MEG-C contains 410 text samples. These contain a total of 664,543 readable words ('tokens'), making up 65,647 lemmata ('types'). More than half the texts belong to the Northern part of the country, while more than a quarter represent Western localizations in LALME, and approximately one fifth represent Eastern ones. Almost half the texts represent what are termed 'documentary texts' in LALME, i.e. legal documents and letters, while the remainder represent a wide range of 'non-documentary' texts, including religious prose and verse, romances, medical texts and various treatises, as well as copies of widespread works such as *Piers Plowman* and the *Canterbury Tales.* Chronologically, the texts range from the early fourteenth century to the early sixteenth; nearly half the texts are dated to the first half of the fifteenth century.

Historical surveys are always restricted to the materials that happen to survive, and therefore do not provide representative, well-balanced samples of the kind that may be compiled using

present-day informants (cf Kretzschmar and Stenroos, forthcoming). The choice of texts in the versions of MEG-C so far has been based on the materials in LALME, which are organised geographically. However, not all LALME localizations are based on the provenance on texts: others are based on the dialectal 'fit' of texts into a matrix representing a postulated dialect continuum. This has resulted in a relatively even coverage of linguistically localized texts over most parts of the country.

The LALME material includes a very wide range of text types. Some of these (for example religious poems such as the *Prick of Conscience*) are distributed fairly evenly over the entire country, while others have a very uneven distribution in terms of the LALME geography: for example, there are few legal documents from the South. In terms of chronology and script type, the distribution is more skewed still: the great majority of LALME texts are dated to the first half of the fifteenth century and are written in an anglicana script.

The eventual coverage of MEG-C, including non-LALME texts, will aim at a representation of materials following clearly defined criteria. In the interim versions, however, even coverage has not been possible due to the largely project-based nature of the work. Large portions of the transcription work have been carried out as part of postgraduate thesis projects or other short-term-funded research projects, and certain areas or text groups have therefore been more fully covered than others. On the whole, the distribution of texts in terms of the LALME geography is skewed towards the northern and western parts of England. This reflects the history of the transcription process. Many of the eastern texts were transcribed at Glasgow in the early stages of the project, and transcription conventions have since evolved, making these early transcriptions more time-consuming to proofread than the later ones; a large number of eastern texts are therefore transcribed but not yet proofread.

The present version of MEG-C (2011.1), as the previous ones, is organised according to the counties within which the texts were mapped in LALME. This organisation has the advantage of being easy to overview and refer to; however, as with LALME itself, it should be taken to reflect actual geographical realities only in as far as the provenance of the texts is given (as it is for most of the documentary texts, which form half the material). A different organisation of the texts, based on non-linguistic variables alone, is currently being developed.

back to top

3. Processing a manuscript text from original/microfilm to a text file

3.1. The selection of tranches

The choice of 3,000 words as sample size will be discussed in the MEG Introduction (see also Stenroos 2007). Normally, a sample consists of two tranches of 1500 words each. In the case of longer texts, these tranches are chosen from different parts of the text, when possible avoiding the first folio or two, as these often show a usage that is untypical of the scribe's normal behaviour (LALME I:15). As these principles were not uniformly applied at the

beginning of the project, some of the earliest transcriptions consist of one or three tranches; however, the total word count should always be ca 3,000. Conventions of word-division and methods of word counting have, however, varied during the project work. The new Concordance version of the corpus files, introduced in MEG-C 2011.1, makes possible accurate and consistent word counting, and word counts are now provided for each text in the corpus. These counts reveal some fluctuation, which will be remedied where possible for future versions.

Whenever possible, the tranches are selected so that they form strictly continuous pieces of text. However, most often they do not correspond to complete works. Given the extremely varying length of the texts localized in LALME, as well as the primary interests of the Middle English Grammar Project (i.e. the study of morphology and phonology), relatively even-sized samples were here considered a more important priority than the completeness of texts.

The Linguistic Profiles in LALME sometimes correspond to more than one scribal text as defined in LALME (I:8). Sometimes, two or more manuscripts that are considered to contain a similar linguistic usage, whether or not produced by the same scribe, have been included under the same Linguistic Profile. In some cases, several documents belonging to the same geographical location and deemed to show the same dialect are also combined in a single Linguistic Profile.

This made sense in the Atlas, where the main objective was to produce a typology of localizable dialectal usages. For the purpose of MEG, however, all such complex profiles are split into separate scribal texts. A scribal text is considered to be either 1) a single work (such as a poem, treatise or sermon) written by a single scribe, or 2) a group of such works that appear consecutively in the same manuscript, with no indication of changing linguistic habits or breaks in copying. Occasionally, we have also accepted two or more documents written by the same scribe at the same time and place as a single scribal text.

Each scribal text included in the Corpus is given a code. For the texts mapped in LALME, this consists of a capital L followed by the LALME Linguistic Profile code, made into a four-digit code by adding initial zeros as necessary (e.g. L0007, L0147, L7340, corresponding to the LALME LPs 7, 147 and 7340 respectively). Where a LALME profile has been split, the separated scribal texts are distinguished by adding a lower-case letter to each code (e.g. L0377a and L0377b, corresponding to the complex LALME LP 377).

In future versions, non-LALME texts will be included in the Corpus (see 1.1. above). These will be distinguished by the use of different capital letters in the code.

back to top

3.2. Transcription routines

Most of the samples have been transcribed from a facsimile reproduction (usually a microfilm printout, photostat, photocopy or digital image); some texts are transcribed from the manuscript itself. A few texts have first been typed in from good diplomatic editions; such transcriptions are always corrected either against the manuscript or a good-quality reproduction.

We hope to check as many texts as possible against the manuscript; this is particularly crucial where the text is difficult to read or the reproduction is of a poor quality. The types of source for each transcription are indicated in the Catalogue of Sources. Each transcription will be proofread at least twice, and every published text has been read by at least two people.

The transcription conventions are based upon those used in the *Linguistic Atlas of Early Middle English*. This should, first of all, ensure compatibility between the two resources. In addition, the LAEME conventions use ASCII characters only, in order to be easily transferable between different platforms; this is a crucial advantage for a long-term corpus project. For the purposes of MEG-C, the conventions have been slightly modified to suit the later material, but the major principles are retained.

The following section gives a detailed description of the transcription conventions. The raw transcriptions are provided as the base text files of the Corpus, and provide the most detailed and accurate record of the physical text in the manuscript. For the purpose of quantitative analysis, a Concordance version of the text files is provided (see 5.1 below), while more 'reader-friendly' versions are provided as html- and pdf-files (see 4 below).

back to top

3.3. Description of the transcription conventions

The transcriptions reproduce the text at what might be called a rich diplomatic level. This includes the following features:

- spelling, distinguishing between 31 letters including the sub-graphemic distinctions between <i/j> and <u/v>, but not other variant forms such as different forms of <r>, single and double compartment <a>, and so on.
- capitalization
- abbreviations and some final flourishes/otiose strokes
- accents over i's.
- punctuation, using the full stop, semicolon, colon and slash for the following types of MS punctuation marks: dot, *punctus elevatus*, colon and virgule.
- word division
- line division, initial large capitals and paraphs
- rubrics/headings
- folio or page references
- some corrections and marginal additions, if plausibly contemporary and helpful for reading the text

back to top

3.3.1. Spelling and abbreviations

The transcription is carried out using only symbols belonging to the basic ASCII set. All ordinary letters are typed in upper case; lower-case letters are used for ME graphs, abbreviations, codes and comments.

The ME graphs <þ, ð, [yogh], æ> are transcribed as the lower case letters or letter combinations <y, d, z, ae> respectively. Of these, only <þ> and <[yogh]> are common in the present corpus.

In many texts, <þ> and <y> are not distinguished; in such cases both are transcribed as <y>, irrespective of what the actual letter form looks like. In the great majority of such texts, the letter form looks like *y*. Where the two letters are distinguished, transcription is strictly according to letter form.

The 'yogh' letter form is used in Middle English for three main functions that are not always straightforward to differentiate between, appearing in substitution sets together with initial <y, yh>, medial/final <gh, h> and mainly (but not exclusively) final <s, z> respectively. Irrespective of function, it is always transcribed as a lower-case <z>. Correspondingly, the zeta-shaped letter form, usually with a cross bar, is transcribed as an upper-case <Z>.

Manuscript capitals are indicated with an asterisk immediately before the letter:

Amen	*AMEN
AMen	*A*MEN

Large or decorated initials that are higher than one line are indicated with two asterisks:

When	**WHEN
WHen	**W*HEN

Abbreviations are transcribed using lower case letters; each abbreviation is transcribed using a conventional 'expanded' spelling in lower case. The aim is to describe the visual form rather than giving a compilatorial interpretation: the ideal is that each formally distinct abbreviation is consistently transcribed the same. In practice, however, it is very difficult to work out what is 'formally distinct' and not; some interpretation is bound to enter the choice. This would still be case even if the abbreviations were rendered with iconic forms or with arbitrary symbols such as $ or }.

The main point is that the 'expansions' are simply ways of indicating abbreviation marks and do not in general involve any assumptions about what these marks 'mean'. Thus, a **macron** is transcribed as a lower case nasal, whether or not this fits our intuitive feel of the linguistic meaning behind. However, this general rule is not carried out *in extremis:* cases where it represents a substantial part of a word (often a proper noun), rather than a single segment, are expanded using lower case letters: e.g. IHesU, CHartRE, IerusaLeM.

Figure 1. The transcription of macrons

Suspension/contraction marks other than macrons are expanded using a set of lower-case expansions, based on the classification of signs of abbreviation by Hector (1966: 30-35). A complete list of expansions, together with references to Hector's classification, is given under 3.4. below.

The most complicated question with regard to transcription conventions has been the treatment of the **final flourishes** that Parkes (1979: xxix) characterized as 'additional strokes which in Latin text would indicate an abbreviation, but which may or may not do so in English'. We take as a starting point Parkes'(1979: xxx) statement that a transcription can afford neither to ignore final flourishes nor to treat them as abbreviations, but should simply record them as final flourishes.

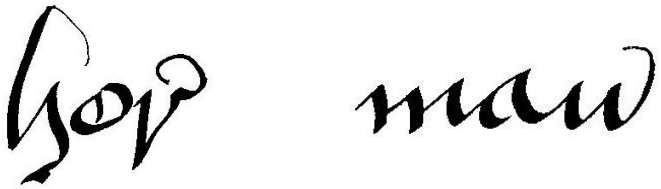
The problem then arises of the definition of a final flourish: at which point does a long end stroke become a flourish? Are cross bars over h's or double l's to be considered 'flourishes' even if they occur completely regularly, or are they part of the regular letter shape (*figura*)?

Such questions can often be answered in a fairly satisfactory way for an individual text; however, for the present purpose it is necessary to follow the same guidelines for every text. Recording everything that could possibly be described as a flourish seemed a hopeless task: some scripts tend to involve something flourish-like in virtually every word, and many flourishes, especially of the cross bar type, seem to be best regarded as part of the *figura*. For example, hands that mark final <ll> with a cross bar generally do so completely consistently; this is borne out both by electronic searches of transcribed texts where the cross bar has been marked and by observation of the usage during the transcription and/or proofreading of at least fifty texts. The variation here tends to be between final single <l> and cross-barred <ll>.

In the end, we have decided to record only such types of flourishes that form part of a continuum either with an abbreviation mark or with a final -e (that is, there are borderline cases that could plausibly be defined as either a flourish or as a fully-formed abbreviation mark or 'e'). This group includes flourishes on word-final minims as well as on (at least) <r>,<g>, <t> and <k>, as well as up-turned flourishes on <d>. The use of other strokes and cross bars will be noted in the Paleographical notes that is planned to eventually form part of the Catalogue of Sources.

The final flourishes fall into two formal categories. The most common are flourishes/endstrokes made without a pen lift, which may be more or less rounded or looped. Such flourishes are here termed **squiggles**. Squiggles are transcribed with a tilde ~ , irrespective of what they 'represent'. Occasionally, this leads to some rather absurd 'readings' such as CUmMYG~ rather than CUmMYnG "coming"

Sometimes, the kind of abbreviations transcribed as <er> or <re> are made without a pen lift and may look identical to squiggles; in such cases the context will usually determine the choice of transcription: OUer THE MOSSE 'across the moor' but THE DAT~ OF THIS INDENTURE 'the date of this indenture', rather than OU~ and DATer.



HER~ MAN~

Figure 2. Examples of squiggles

Another type of flourishes are made without a pen lift, but involving a change of direction, so that they end up as a combination of a squiggle and a macron. Their functions also appear to combine those of the squiggle and the macron, in that they seem to commonly represent both a nasal and a potential final *-e*. Such flourishes are here termed **squigrons,** and have been transcribed as @.

Figure 3. Example of a squigron

The transcription does not aim to preserve graphetic detail: different variant forms for the same letter (e.g. double/single compartment <a>) are not distinguished. As an exception to this general rule, accents over <i> are in the base files indicated with a % sign following the letter. This is because such accents frequently determine the readings of minim clusters; thus, their presence or absence may be of importance in evaluating the justification for our reading. On the other hand, dots over <y> are not included. While they sometimes help to disambiguate <y> and <þ>, they add much less information for the interpretation of the text, and including them has not been deemed economic.

back to top

3.3.2. Punctuation

Punctuation is indicated using the following signs as appropriate: < / ; : . >. A gap is left between the last letter and the punctuation mark: HE CAME . AND SAW . This convention is most faithful to the usual manuscript usage, and facilitates the process of entering the transcribed texts into the database for annotation, as punctuation marks need to be separated from words. Earlier, it also helped to distinguish between scribal and editorial punctuation during the first stage of work, as editions were sometimes used for producing first drafts of transcriptions.

back to top

3.3.3. Word and line division

Manuscript word division is retained. We do not, however, measure the gaps between words: if what we think about as two words are not very obviously written together, then they are deemed to be written apart and transcribed accordingly. However, in clear departures from present-day word-division, the following conventions are used:

Where two words (as defined by the headwords of the *Oxford English Dictionary*) are clearly written together, they are transcribed together, with a + sign indicating that we are dealing with what might be analysed as two separate words (e.g. A+MAN “a man”). Conversely, when what we (and the *OED*) would consider a single word is divided into two parts, these are combined in the transcription with a hyphen: so WHER-FORE “wherefore”. It should be noted that, while the uses of + and - do preserve the manuscript reality, they also impose an interpretation on the text (see 3.5. below); using them at all is a purely compilatorial choice, designed to make the next stage of analysis easier.

The text is transcribed line for line, with manuscript line division marked by the Return key. Line numbering can thus be added to the transcriptions when wished. In the base text files, word division at the end of lines is marked by adding # to the end of the first half, e.g.

HAP#
PY

If the scribe has marked the division with a hyphen (usually a double diagonal stroke), this is indicated with a = symbol before the hash:

HAP=#
PY

The word divisions across lines are retained only in the Base files; in the Readable version they have been silently removed, while the Concordance version marks the divisions but brings the words together to make them count as single units.

back to top

3.3.4. Foliation

Foliation (alternatively pagination) is indicated throughout. The beginning of a new folio is indicated within angled brackets in the following format: <fol. 8r> (alternatively, <p. 8r> for a paginated text). If the transcription does not begin at the top of a page, line number is indicated as well, in the format <fol. 8r><line 10>. Columns of text are indicated with lower case <a, b...>. Thus, a folio with two columns of text on each page will consist of the following four sections: <8ra>, <8rb>, <8va>, <8vb>.

back to top

3.3.5. Coding for layout, corrections and commentary

A set of codes placed within angled brackets are used to indicate specific layout features, corrections and additions, as well as other kinds of commentary:

Rubrics and headings are marked by inserting the following codes before and after the text: <rub>...</rub>. Underlining is marked with the codes <und>...</und>. Expunction or crossing out is marked with the codes <exp>...</exp> and <cro>...</cro> respectively. Partially rubbed-out text is marked with <rbd>...</rbd>.

Added/inserted text may be marked in four ways, depending on where it has been added. Whether added by the scribe himself or a later corrector, text is most commonly inserted above the line or in the margin; such insertions are marked with ^{...} and <mrg>... </mrg> respectively. Occasionally, an addition is made in an existing gap within the text or over a rubbed-out section; in such cases, it is marked with <add>...</add>. If the addition is marked with a caret, the code <ct> is used. Thus, the sequence W<ct>^HAT in the base text indicates a spelling *wat*, with <h> added above the line and a caret between W and A.

The code _{...} indicates the continuation of a line at the right hand end of the following line. In general, insertions do not appear in this position; if they do, they are marked <add>...</add>.

Additions are generally transcribed if they are are considered to be at least potentially contemporary and/or important for understanding the text. Often it is impossible to tell whether they were carried out by the same scribe or not. Therefore, any text marked with the <sup>, <mrg> or <add> codes should be excluded when studying the language of a particular scribal text.

Latin words or passages within the text are marked with the codes <lat></lat>, and are usually not transcribed. In the Base version, the codes are repeated for each line in order to preserve the line count of the text.

Illegible letters or passages are marked with the code <ill>...</ill>. The approximate amount of text missing is indicated within angled brackets: <ill><c. 2-3 words></ill>. Sometimes, the last portion of a line may be invisible because it disappears into binding, or it may have disappeared if the pages have been cropped; in such cases, a descriptive comment is placed within angled brackets, e.g. <binding>, <cropped>.

Finally, any comments may be placed inside angled brackets, and written in ordinary lower case: e.g. <writing slightly smudgy here>, WUN <? four minims>. Such comments appear in the Base text files but are removed from the Reading version unless deemed crucial for the reading of the text; all comments are removed from the Concordance version but an exclamation mark ! after a word indicates that the Base files should be consulted for more information.

back to top

3.4 List of symbols used in the transcription

The following list of symbols summarizes the transcription conventions used in the Base files; for a description of their use, see 3.3. above. Abbreviations are defined according to the classification by Hector (1966: 30-35) using his classification numbers. Non-alphabetic symbols are listed first, then letters and finally codes enclosed in angled brackets.

< >	enclose anything that is not to be read as part of the transcription, such as codes and comments
; :	*punctus elevatus*
.	*punctus*
/	*virgule*
&	any symbol used for ‘and’
~	squiggle (= a word-final flourish that may either be functionally equivalent to <e> or otiose)
@	squigron (= a squiggle combined with a macron, i.e. a flourish that involves a change of direction)
%	acute accent or ‘dot’ over <i>
\	defines following letter as a superscript one (used only for the systematic use of superscript as in $þ^{t}$ ‘that’; not used for corrections or additions above line)
#	signals word division across the line

=	word division marker in the manuscript (always placed before #)
-	gap between two words that would correspond to a single word in Present-day English usage (e.g. *to-geder* 'together')
+	assumed boundary between two words written together in the manuscript but corresponding to two separate words in Present-day English usage, e.g. *a+man* 'a man'
*	defines the following letter as a capital
**	defines the following letter as a large initial capital extending over more than one line
a, ua, ra	expanded abbreviation (derived from superscript *a*, cf Hector 1966: 34-35)
ae	the letter <æ>, 'ash'
con, com	expanded abbreviation (Hector 6)
d	the letter <ð>, 'eth'
er, ar, re	expanded abbreviation (Hector 3)
es	expanded abbreviation (Hector 9)
ir, ri	expanded abbreviation (derived from superscript *i*, cf Hector 1966: 34-35)
n, m	macron (Hector 2)
Per, Par	expanded abbreviation (cf Hector 1966: 34)
Pro	expanded abbreviation (cf Hector 1966: 34)
ur	expanded abbreviation (Hector 4)
us	expanded abbreviation (Hector 5)
y	the letter <þ>, 'thorn'
z	the letter <3>, 'yogh' or the similar-shaped variant form of <z>
Z	the zeta-shaped variant form of <z>
<add></add>	enclose text added on the same line, in gap or over erasure, either by the same scribe or by a later corrector (clearly post-medieval corrections are ignored)
<brd></brd>	indicate the convention in some verse texts (particularly carols) to position single metrical lines, following rhyming couplets, and rhyming with each other, at the right hand side of the couplets
<cro></cro>	enclose text that has been crossed over for deletion
<ct>	caret
<exp></exp>	enclose text that has been expuncted for deletion
<ill></ill>	enclose illegible text (approximate amount of text indicated in diagonal brackets between the codes)
<lat></lat>	enclose text in Latin
<mrg></mrg>	enclose text added in the margin either by the same scribe or by a later corrector (clearly post-medieval corrections are ignored); placed at caret position when marked
<pph>	paraph
<rbd></rbd>	enclose text that has been rubbed out/erased; if illegible, the approximate amount of text is indicated in diagonal brackets between the codes
<rub></rub>	enclose text marked as a heading/rubric or strongly emphasized by means of script type/size or colour
	enclose text continuing below the line, usually at end of the following line
	enclose text added above the line either by the same scribe or by a later corrector (clearly post-medieval corrections are ignored); placed at caret position when marked. Used only for corrections or additions above the line, not for the systematic use of superscript, as in *þ*^{t} 'that'.

back to top

3.5. Editorial decisions and interpretation in the Base files

On the whole, the transcription aims to record what is visible in the manuscript, rather than giving editorial interpretations. However, any transcription will inescapably involve an element of interpretation. The users of the present Corpus should in particular be aware of the following compilatorial choices in the Base files:

Firstly, the uses of #, - and + entail compilatorial interpretations of word division. A user who does not wish to be influenced by these may download the text and make the following substitutions: zero for # and +, and space bar for - .

The choice between \ and when marking superscript letters is based on the transcriber's understanding of the distinction between the systematic use of superscript letters as abbreviations (e.g. w^t, $þ^t$, $þ^u$) and the unsystematic insertion of letters above line for the purpose of correction and addition. The latter may be added by a later correctors, and it is often impossible to tell whether this is the case or not, especially from a microfilm reproduction. The compilers have therefore not attempted to distinguish between additions by the same or another scribe in the transcription, with the exception of clearly post-medieval additions, which are ignored. In general, it is therefore advisable to treat with caution all text that appears between the codes , <add></add> and <mrg></mrg>, and not to take for granted that they represent the same scribal usage as the rest of the text.

The reading of minims often entails interpretation. As accents over <i> are recorded in the transcription, the user will be able to determine in which cases they have clarified the reading. Where such accents are absent and the script makes no distinction between <u> and <n>, a sequence of six minims transcribed as MIN, NIM or NUN is based on the transcriber's judgment of what fits the context best. The same applies, in many texts, to the choice between <st> and <sc>.

Finally, in some texts, squiggles (~) and *er/re* abbreviations (Hector 3) may also look identical, and the choice is then a matter of interpretation. Similarly, the choice between superscript <i> and the *ir/ri* abbreviation is often compilatorial: the two are historically identical, and often (if not always) identical in form. As far as this last distinction goes, the user who does not wish to be influenced by our choices may replace lower-case <ir> and <ri> with <\I> in their downloaded copy of the Base Corpus files.

Some of the compilatorial choices have been necessary from the point of view of the use of the data: it is, for example, important to distinguish between such superscript letters that are part of the scribe's spelling system and ones that represent additions that may have been carried out by another scribe. Others would have been possible to avoid, and may be removed by the user, using substitutions such as those suggested above. The possibility of indicating minims by some neutral sign, rather than interpreting them as specific letters, was discussed by the team, but was then abandoned for two main reasons: the shortage of suitable ASCII characters and the distinct loss of readability, not only from the point of view of users outside the project but also the co-workers entering the text into a database.

back to top

4. Different flavours of the corpus

The Middle English Grammar Corpus is published in three different flavours.

a) The first flavour is called **MEG-C Base.** The files in MEG-C Base are in UTF-8 format, and the text is presented as has been described under 3 above. MEG-C Base contains the transcriptions that reflect manuscript reality most closely, as well as most of the information and the annotation added by the compilers. Thus this version is the one that the users of MEG-C should consult when in need of more information. The files of this version can either be viewed on-line or downloaded as a .zip archive.

b) The second flavour of the corpus, **MEG-C Readable,** represents the texts as .html files and .pdf files. This version is meant for easy browsing and reading the pages on screen, or for printing out the text for reading. The differences between MEG-C Base and MEG-C Readable are as follows:

- In MEG-C Readable, the default case is lower case. Capital letters are represented in CAPS, making the coding for them unnecessary, and abbreviations are expanded in italics.
- Words divided from a line to a new line have been joined silently.
- All scribal and compilatorial coding has been deleted, so that paraphs, underlining, superscript, deletion etc. are represented iconically.
- Compilatorial comments have been kept to the minimum

The Readable files may be viewed on-line, and they are also available in a .zip archive intended for downloading.

c) The third flavour of the corpus**, MEG-C Concordance,** is a version of the text files that has been modified for efficient use with a concordancer or with other corpus software. It is described in detail in the following section, with suggestions for its use together with the concordancing programme AntConc. The files may be downloaded as a .zip archive.

back to top

5. Searches and recommended software

There is no search function implemented on the web site. The recommendation is that the text files are downloaded and then used with the text processing or corpus software of one's choice. The downloadable files of the text files are UTF-8 encoded and the end-of-line coding follows the UNIX format. However, the files are ASCII compatible: we use only the first 127 characters of the UTF-8 set, and those are identical with the first 127 characters in the basic ASCII set. Therefore the text files are suitable for any concordancing program that can digest ASCII, such as AntConc or WordSmith. As the transcription distinguishes between upper and lower case letters for several purposes, it is important that the chosen program support case sensitivity.

5.1. The MEG-C Concordance version

The 2011.1 version of MEG-C introduces a version of the corpus that is intended specifically for quantitative analysis using a concordancer or other corpus sofware. This version, referred to as MEG-C Concordance, or the Concordance version, lacks the two features of the Base version that make it less suitable for concordancing: large amounts of coding and comments, and word division at line breaks. It is designed with a particular concordancing programme, AntConc 3.2.1. , in mind, but should suit most case-sensitive programmes as long as the definitions of word characters (characters that are read by the programme as forming part of words) are adjusted as necessary.

The main principle has been to distinguish strictly between word characters and non-word characters, so that everything that is not meant to be read as a word is either removed or turned into non-word characters. The choice of characters has been based on the requirements of AntConc 3.2.1. The main differences between *MEG-C Base* and *MEG-C Concordance* are as follows:

- All headers, tranche headings and folio/page/line numbers have been removed
- All coding that does not directly affect the readings has been removed, including <und>, <sub> and <brd> are removed, as well as paraphs and carets. The '%' sign indicating accents/dots over i's is also removed.
- The <lat>...</lat> sequences indicating Latin text have been replaced with '£'
- Any other string of content that does not represent readable words (illegible or deleted words; symbols) is replaced with the symbol '¤'
- Partially legible words are only retained where there are only one or two missing letters and the reading is otherwise clear; in such cases, missing letters are substitutd by a comma, as in BE,ORE. In other cases, the entire word, including the legible letters, is replaced with '¤'. Letters that are marked for deletion, such as an expuncted I in ABOIUE 'above', are removed, but an exclamation mark is added after the word to direct the user to the Base version for information.
- Codes that indicate insertions (<sup>, <mrg> and <add> are replaced by curly brackets '{ }'. All words within such codes are bracketed individually.
- Line division is marked with '[' but divided words are written in one, followed by a line break.
- Comments are removed and exclamation marks added to direct the user to the Base version for information.
- Certain characters that are used as wildcard characters by AntConc are replaced by others; for example, both squiggles and squigrons appear as '~', freeing the '@' sign for wildcard use, and capitalisation is indicated by ':' freeing the asterisk for wildcard use. For this purpose, it has been necessary to remove the distinction between ':' and ';' which has been deemed less important (and in any case often difficult to make) than the use of wildcards. Finally, the '+' used for separating words that are written together is replaced by '_', freeing the '+' sign for wildcard use.

The following conventions are used in the Concordance files (NB that the lower case 'expansions' of the abbreviations are exactly the same as in the Base files):

< ?>	occurs very occasionally to indicate a particularly problematic reading that the compilers are still considering; not meant to be a regular feature
;	*punctus elevatus* or colon in the manuscript
.	*punctus*
/	*virgule*
&	any symbol used for 'and'
~	squiggle or squigron (see section 3.3.1)
\	defines following letter as a superscript one (used only for the systematic use of superscript as in *þ*^{t} 'that'; not used for corrections or additions above line)
[	signals word division across the line; the two parts are written together with line shift following: BI[FORE
=	word division marker in the manuscript: BI=[FORE
-	gap between two words that would correspond to a single word in Present-day English usage (e.g. *to-geder* 'together')
_	assumed boundary between two words written together in the manuscript but corresponding to two separate words in Present-day English usage, e.g. *a+man* 'a man'
:	defines the following letter as a capital
::	defines the following letter as a large initial capital extending over more than one line
{ }	enclose text added above line, in the margin or on the same line, in gap or over erasure, either by the same scribe or by a later corrector (clearly post-medieval corrections are ignored). Each word in a sequence of inserted words is bracketed, making it possible to identify them all.
¤	indicates the presence of a string in the text that is not a readable word; for example a symbol or a wholly or partly illegible word, or a crossed-out word.
£	indicates a string in Latin. Repetition of complete lines of Latin is not marked.
!	indicates that the user should check the reading in the Base files before using it as a research finding. It often means that a reading is uncertain, but it may also mean that the word contains one or more expuncted letters, that it has been tampered with, or the like.

For best use of the Concordance files, the programme used should recognize the following characters as 'word characters' or 'token classes', i.e. read them as part of words:

abcdefghijklmnopqrstuvwxyzABCDEFGHIJKLMNOPQRSTUVWXYZ1234567890-&[=~\{},

The following characters should NOT be read as part of words:

<>?;./_:¤£!

5.2. Using the Concordance files with AntConc 3.2.1

AntConc is freeware, and full instructions for use are provided on the designer's website. The following is a list of simple steps to get started, but the instructions should be referred to in order to make the best use of the programme.

1.Open the AntConc main window.

2. Go to Global Settings .

3. On the list on the left hand side, choose Token (Word) Definitions. On the right hand side, find User-defined Token Classes. Copy and paste the string of 'word characters' into the available field:

> abcdefghijklmnopqrstuvwxyzABCDEFGHIJKLMNOPQRSTUVWXYZ1234567890-&[=~\{},

4. Then click the box 'Use Edited Definition' and finally click 'Apply'. You are now ready to start.

5. Click Open Directory or Open File, depending on how many files you want to search at one time. You may wish to group the files into different folders in order to study specific subgroups at a time. Make sure to close files that you do not wish to include in a search. If you include all 410 files at one time, the searches may be somewhat slow. Use the 'Word List' function to produce a list of all the words sorted according to frequency or according to first or second letter. Click on a particular word for a KWIC view, and again on the particular word for context within the text.

back to top

6. Feedback

We welcome feedback on any issues ranging from manuscript readings to more technical aspects of digital text representation. If you have a question, a request or a comment, please do not hesitate to contact Merja Stenroos: merja dot stenroos at uis dot no or Jacob Thaisen: jacob dot thaisen at uis dot no

back to top

7. Updates

Each updated version will receive a new version number. Older versions of the corpus will be stored as .zip archives, so that they will still be available after the changes. Information about updates will be posted in the News section of the MEG website.

back to top

References

Hector, L.C. (1966), *The handwriting of English documents.* 2nd edn. London: Edward Arnold.

McIntosh, A., M. L. Samuels & M. Benskin, with M. Laing & K. Williamson (1986). *A Linguistic Atlas of Late Mediaeval English.* 4 vols. Aberdeen: University Press.

Jordan, R (1968), *Handbuch der mittelenglischen Grammatik: Lautlehre,* 3. Auflage, Heidelberg: Carl Winter / Universitätsverlag (first publ. 1929).

Kretzschmar , W.A. and M. Stenroos (forthcoming), 'Atlases and Surveys as evidence' in T. Nevalainen & E. Traugott (eds), *Rethinking the History of English.* Oxford: Oxford University Press.

Parkes, M (1979), *English cursive book hands, 1250-1500. London: Scolar Press.*

Stenroos, Merja (2007), 'Sampling and annotation in the Middle English Grammar Project' in A. Meurman-Solin and A. Nurmi (eds), *Annotating Variation and Change*. Helsinki: University of Helsinki.

.

back to top