

MEG-C

Corpus Manual - version 2009.1

Merja Stenroos and Martti Mäkinen
December 2009



[1. Preliminaries](#)

[1.1. Introduction](#)

[1.2. The purpose of this manual](#)

[1.3. Notes on terminology](#)

[2. Principles of compilation](#)

[2.1. The eventual scope of the corpus](#)

[2.2. The selection of texts for version 1.0](#)

[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](#)

[6. Feedback](#)

[7. Updates](#)

[References](#)

1. Preliminaries

1.1. Introduction

The *Middle English Grammar Corpus* (MEG-C) is a text corpus consisting of samples of the texts localised in the *Linguistic Atlas of Late Mediaeval English* (McIntosh, Samuels and Benskin 1986, henceforth LALME). Shorter texts are included in their entirety and longer ones in 3000-word samples. More than a thousand texts were localised in LALME; the corpus will eventually include samples of all those texts that the project team will be able to access. Apart from the LALME texts, it is planned that the corpus will in the future also

contain subcorpora of Early Middle English texts (1100-1300) as well as of late mediaeval texts not included in LALME. The LALME texts are, however, the first priority.

The corpus will form the main source material for further work within the [Middle English Grammar Project](#). All the text samples are entered into a database together with information about extralinguistic variables such as date, genre, script type, etc.; 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 will be 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 MEG-C version 1.0 was launched in April 2008. The corpus will be updated regularly as more texts 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 will also be updated for each version.

[back to top](#)

1.2. The purpose of this manual

This manual describes the sampling, transcription conventions and presentation of the corpus. 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 3000-word **samples**, which normally consist of two 1500-word **tranches**.

The word **text** may here be applied to either 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 eventual 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 MEG-C is, in the first instance, defined by a single external criterion: inclusion as a localized text in LALME. If 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 LALME-based corpus, when finished, 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).

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](#) at the University of Edinburgh.

The chronological scope of the corpus is the same as that of the main LALME material, that is, ca 1325-1500. During the compilation of LALME, a separate survey of earlier materials was not yet envisaged, and thus a small group of thirteenth-century texts was 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 the present LALME-based corpus, but will, it is hoped, eventually form part of a subcorpus of Early Middle English texts. Text materials from the earlier period (1150-1325) are now available in the [Linguistic Atlas of Early Middle English](#) at the University of Edinburgh.

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 be able to include every single one of the texts defined above. 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 may simply be difficult or impossible to access. The main principle of compilation for the present

corpus must thus be a relatively flexible one: the corpus seeks to represent as large a proportion as possible of the texts localized in LALME, excluding the Scottish and Early Middle English texts.

Within each version of the corpus, it would be desirable to present as full a range as possible when it comes to the geographical and chronological distribution of texts, as well as the representation of text genres and script types. The geographical coverage of LALME is in itself far from even: some areas simply provide more material than others. The distribution in terms of genres is similarly uneven. However, in terms of both geography and genre distribution, there is a very wide range, with large groups of texts from many areas and genres. 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 interim versions of the corpus should, ideally, reflect these distributions in LALME, if possible erring on the side of a more even overall coverage (for example, including a relatively high proportion of non-anglicana scripts). However, as the selection of texts within the first versions of the corpus is ultimately dependent on the availability of texts and on the practicalities involved in their transcription and proofreading, it may fall somewhat short of this ideal.

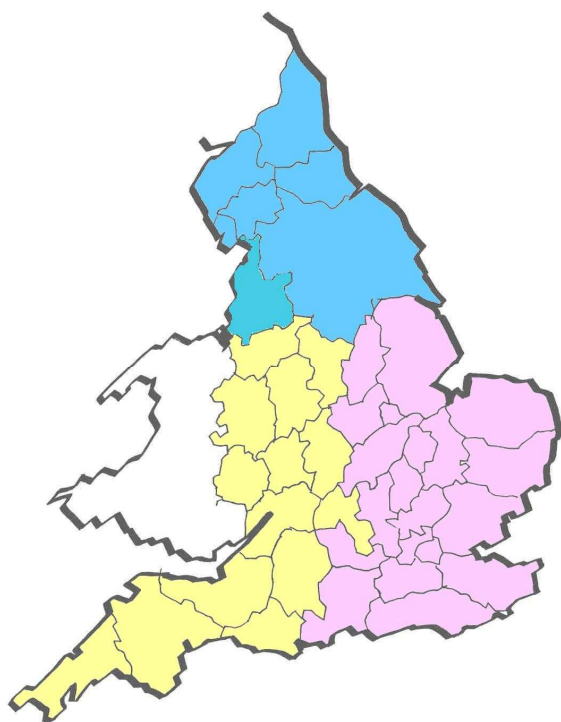
[back to top](#)

2.2. The selection of texts for version 1.0

The present version of the corpus contains 320 text samples. Altogether, these contain ca 450 000 words.

The geographical distribution of texts is skewed towards the Northern and Western parts of England. This reflects the history of the transcription process. When the project work began in 1998, it was agreed to divide the geographical area into three main regions: the East, West and North (see [map](#) below). The Glasgow team took responsibility for the Eastern part, and, as most of the early transcription work was carried out at Glasgow, the majority of the first texts to be transcribed were Eastern ones. With project funding at Stavanger, large-scale transcription of the Western and Northern materials began in 2006. At this point, the transcription conventions had evolved considerably. Proofreading the later transcriptions has been easier and faster than proofreading the early ones, and facsimile reproductions against which to proofread have also been more readily available to the Stavanger team. As a consequence, the corpus so far contains a relatively low proportion of Eastern texts. For the next version, it will be a priority to rectify this geographical skewing.

Version 2009.1 contains a somewhat higher proportion of legal documents than the LALME material at large. This is partly for reasons of availability, but it should also be useful for the purpose of comparing what may be considered the two main groups of late-medieval texts, viz. legal documents and ‘literary texts’ (cf LALME I: 39). Otherwise, the genre distribution follows broadly that in LALME. There has been no attempt to select texts with any particular chronological distribution pattern in mind: the texts range from early fourteenth-century ones to texts dated to around 1500; nearly half the texts are dated to the first half of the fifteenth century.



Map 1. The regional division of the corpus material: East, West and North

[back to top](#)

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

3.1. The selection of tranches

The choice of 3000 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). However, these principles were not uniformly applied at the beginning of the project. Thus, some of the earliest transcriptions consist of one or three tranches; however, the total word count is always ca 3000.

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), 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 localized in LALME, this consists 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).

Eventually, it is planned to include non-LALME texts in the Corpus ([see 1.1. above](#)). These will be distinguished by the use of different capital letters in the code. For example, texts localized in *The Linguistic Atlas of Early Middle English* (LAEME) could appear with codes such as E0003 (corresponding to text #3 in LAEME).

[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 ASCII restriction and 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, while more ‘reader-friendly’ versions are given 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* (with or without a long top stroke) and virgule.
- word division
- line division, initial large capitals and paraps
- 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, [wynn]> are transcribed as the lower case letters or letter combinations <y, d, z, ae, w> 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; in the few cases where it looks more like þ, this will be noted in the paleographical notes provided in the Catalogue of Sources (available from Version 1.1.). 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 always 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:

W hen	**WHEN
W Hen	**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 high 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 CUMYg~ rather than CUMYnG "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 determine the choice of transcription.

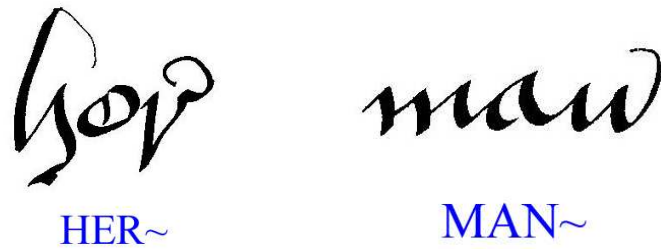


Figure 2. Examples of squiggles

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 are marked only in the Base files; in the Html version they have been silently removed. It should be noted that a concordancer-based analysis of the Base files of prose texts will need to take the word divisions into account; where this is problematic, the Html files should be used instead.

[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.

Finally, the code _{...} indicates text added below the line. In practice, insertions seldom appear in this position; however, the code is also used to mark the continuation of a line at the right hand end of the following line.

Additions are generally transcribed if they 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. 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 <or is it WIM?>. Such comments appear in the base text files but are removed from the other versions of the Corpus unless deemed crucial for the reading of the text.

[back to top](#)

3.4 List of symbols used in the transcription

The following list of symbols summarizes the transcription conventions used in the corpus; 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
; :	<i>punctus elevatus</i>
.	<i>punctus</i>
/	<i>virgule</i>
&	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 <i>b^t</i> ‘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. <i>to-geder</i> ‘together’)
+	assumed boundary between two words written together in the manuscript but corresponding to two separate words in Present-day English usage, e.g. <i>a+man</i> ‘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 <i>a</i> , 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>i</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)
<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 <i>b^t</i> ‘that’.

[back to top](#)

3.5. Editorial decisions and interpretation in the Corpus

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 user of the present Corpus should in particular be aware of the following compilatorial choices:

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*, *b^t*, *b^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 will have to be 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 three reasons: the shortage of suitable ASCII characters, the work involved in making the changes to all the transcriptions made so far (nearly 500 at that stage of the work) 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 two 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 Html](#) represents the texts as .html files. This version is meant for easy browsing and reading the pages on screen. The differences between *MEG-C Base* and *MEG-C Html* are as follows:

- In *MEG-C Html*, 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 paraps, underlining, superscript, deletion etc. are represented iconically.
- Compilatorial comments have been kept to the minimum

In Version 1.1., there will be links to Catalogue entries from the corpus file headers.

MEG-C Html is also available as .pdf files, the links to which are found beneath each corresponding .html link. These files can be viewed on-line, and they are also available in a .zip archive intended for downloading.

[back to top](#)

5. Searches and recommended software

There is no search function implemented on the web site yet. The recommendation is that the text files are downloaded and then used with text processing or corpus software of one's choice. The downloadable files of MEG-C Base 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, e.g. such as AntConc or WordSmith. As the transcription methods distinguish between upper and lower case letters for several purposes, we advise that the chosen program support case sensitivity.

[back to top](#)

6. Feedback

We will 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](mailto:merja.stenroos@uis.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.

Between the updates, no new texts are added to the corpus. Errors will, however, be corrected as they are found; all such corrections will be noted in a List of Corrections posted on the corpus page.

[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).

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](#)