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 Concordancer 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 Concordancer 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:

<>?;./_:¤£!

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.