ARCHER 3.2: guidelines for new texts

Introduction

Teams adding new texts are asked to adhere to the following conventions on filenames, headers, caret brackets, hyphens, dashes and special characters: easy to follow at file creation but laborious to impose later. These conventions were retrospectively applied as far as possible to ARCHER 3.1.

Filenames and file dates

All filenames are consistently 8+dot+3 characters (always lower case), according to the formula **nnnnabcd.gpv**, where

- **nnnn** = year (if necessary with x's for uncertain or more than one year)
- **abcd** = abbreviation of author's name (usually first four letters), padded out with hyphens if too short, and occasionally with d = numeral if there is more than one file from the same year
- **g** = genre, according to formula a = advertising, d = drama, f = fiction, h = sermons, j = journals, l = legal, m = medicine, n = news, s = science, x = letters, y = diaries
  [NB. New genres 'advertising' and 'diaries'.]
- **p** = period, according to formula 0 = pre-1600, 1= 1600-49, 2 = 1650-99, 3 = 1700-49, 4 = 1750-99, 5 = 1800-49, 6 = 1850-99, 7 = 1900-49, 8 = 1950-99, 9 = post-2000
  [NB. Period 1 files will be needed for the first time in the ARCHER 3.2 release, and periods 0 and 9 are not needed at present]
- **v** = variety, according to formula b = British, a = American

The same author should always have the same four-character abbreviation, unless there is another extract from the same year, as with 1951fknr and 1951fkn2 – and a numeral should then replace the *last* letter. (NB. the reverse implication does not hold, e.g. "whit" represents numerous different authors.)

All files in a given release will be date-stamped with a consistent date and time in Manchester prior to release to allow subsequent changes to show up easily in directory listings, but compilers should send files with dates as they are.

Headers

The current filename is the first item in the header. An accurate word count is the second item. Bibliographic info follows. All header material is now placed within carets. It may be necessary to restrict use of certain punctuation characters in headers: watch for updates.

If the file is not completely new but dates back from before ARCHER 3.1, the previous filename may also be shown later in the header, sometimes with even older variants retained from earlier incarnations.

Brackets

We urge very strongly that for newly transcribed texts, **all and only material** inserted by corpus compilers or editors should be placed within caret brackets.

Word counts

The second item in every header is an accurate word count of everything in the file which is not enclosed in caret brackets. That means that no header material is counted, nor comments

from the compilers enclosed in caret brackets. It also means that stage directions in some texts are not counted. Speaker names in curly brackets – used in some drama texts – are not counted either. Material within square or round brackets is counted if not also enclosed by carets. Hyphenated words are counted as one item, as are all items other than punctuation surrounded by white space. The Perl script which we use to count words is on the **documentation** page of the ARCHER website, or we can standardise word counts in Manchester when collating new files.

We suggest aiming for a typical word count per text of 2,000-2,500 words: although short by modern corpus standards, it's close to the existing average in ARCHER (typically 2,000 outside Fiction) and so won't make new genres unbalanced.

Hyphens and dashes

All dashes must appear as a double hyphen -- like that -- with one space to both left and right unless at line-end or beginning. (NB. word processors often automatically change such coding into an em-dash.) This clearly differentiates punctuation from the hyphen, which is single and does not have white space around it. Special characters representing em- or en-dashes should not be used. Hyphens should not appear as the last character of any line: the second element is taken back from the next line and the whole thing made into either a single unbroken word or a hyphenated word.

Special characters

File transmission between different operating systems (e.g. the original DOS and various versions of Windows and Mac OS), and even sometimes the use of editors or other programs, can corrupt special (non-ASCII) characters like *ä, £,* ° [a-umlaut, pound, degree sign]. Please keep a note of characters used, and by all means send a sample file to Manchester to see if they survive the various transmission processes. The file **non-ASCII_chars_3-1.txt** released with ARCHER 3.1 may serve as a reference. We will almost certainly move towards so-called HTML entities for special characters in the standard distribution, allowing users to edit locally to their desired appearance if they wish (and providing a simple conversion routine).

DD and NYB (drawing on a document distributed with ARCHER 3.1 in 2006)

4 September 2009