ARCHER 3.1, July 2006: problems and edits

Source files used for version 3.1

Two ARCHER-1 versions were combined: files from ARCHER-1a (German universities mainly, where many files, esp. fiction, had apparently been thoroughly corrected at some point); and from ARCHER-1b the science text 1825barl.s5b was filled with text (empty in ARCHER-1a), and 1674ano1.s2b was included (missing in -1a).

Future additions

Teams adding new texts are asked to adhere to the present conventions on filenames, headers, caret brackets, hyphens, dashes and special characters: easy to follow at file creation but laborious to impose later.

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 longstanding formula d = drama, f = fiction, h = sermons, j = journal or diaries, l = legal, m = medicine, n = news, s = science, x = letters
- **p** = period, according to new 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
  [period 1 files are being held over until the ARCHER 3.2 release, and periods 0 and 9 are not needed at present]
- **v** = variety, according to revised 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 this release have been date-stamped as 31 July 2006 03:10 to allow subsequent changes to show up easily in directory listings.

Headers

The current filename is the first item in the header. An accurate word count is the second item. The previous filename is also shown later in the header, sometimes with even older variants retained from earlier incarnations. Bibliographic info follows. In this release, all header material is now placed within carets, and unbracketed lines with a filename preceded by asterisks have been altered. Many headers were corrected when the texts were being verified at Heidelberg, and the format has been partly regularised.

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, <>. Outside the headers, existing caret brackets have been retained as they were, however. In some texts they enclose original spellings as opposed to normalised spellings or explanations – the logical opposite of

what we would suggest for a historical corpus – and carets also enclose original stage directions in some drama texts. Unpaired brackets have been corrected for caret <>, curly {} and square [], but not for round () brackets.

Hyphens and dashes

All dashes appear as a double hyphen with one space to both left and right -- like that -- 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 no longer appear as the last character of any line: the second element has been 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]. We have attempted to correct these characters. Wherever possible the present release of ARCHER 3.1 will display such characters correctly if the text encoding is set to Western (Windows Latin 1). One of the accompanying text files lists all such characters, so that users who find them displaying wrongly in an ARCHER text can use that list to interpret them or even change them appropriately on their own systems. The list may also serve as a reference for those transcribing new texts.

There is a version of the ARCHER 3.1 files (ARCHER_3-1_ascii.zip) where such characters are stored not in Windows form but as HTML entities (e.g. *ä £ °*), which will make it easier in the future to move ARCHER towards XML or Unicode. The default release of 3.1 uses bracketed pseudo-words instead of the two mathematical operators which look like caret brackets, and likewise it has bracketed letter-names instead of Greek letters, thus *[less-than], [greater-than], [alpha], [beta],* etc., while the alternative version has HTML entities in both cases: *< > α β* etc.

Word counts

The second item in every header is now 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 – see above – 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. There is full list of every "word" in the corpus, in descending order of frequency, on the CD, plus the Perl script which did the count. The document **numbers_of_files_&_words_3-1.doc** adds up the new figures separately for British and American varieties, with subtotals for each period and each genre. The grand total is **1,789,309** words.

Versions of ARCHER 3.1

Allowing multiple versions always runs the risk of future confusion. Nevertheless we are putting several slightly different versions on the same CD to suit different users or uses:

- The basic version: 955 separate text files, in the Windows Latin 1 character set, placed in a single folder and dated 31 July 2006, 03:10.

- Exactly the same files but split into 80 separate folders or directories, one for each genre-period-variety combination. Files are stored in a ZIP archive which can be extracted with or without its folder structure intact.
- An alternative version – 169 of the 955 files differ – containing HTML-like entities such as *ä é £* instead of single characters for accented letters, pound, etc. This is otherwise identical to the main set and the word counts have not been altered for it. These files are date-stamped 23 July 2006, 03:10 and are stored in a ZIP archive.
- The whole of ARCHER 3.1 in a one-file version where each (very long) line is a single file of the basic version, but with that file's header reduced to filename only. The file is compressed in a ZIP archive.

Other problems to be noted

- A number of British and Irish texts had silently been given American spelling when first transcribed (*recognizes, quarreling, defense, unsavory, behavior, practice* vb, *offense, color, neighbors, traveling, candor, parlor, fulfill, favor, offense, favorite, honor*, etc.): not corrected.
- Place-names were sometimes artificially hyphenated: *Black-River, Buckingham-House*, etc.: not corrected.
- Some files finish (or less often, start) in mid-sentence: not corrected.
- Spacing was sometimes inserted between quotation mark and quotation, between text and punctuation, between elements of contracted forms like *I'll*. This layout was only intermittently applied, even within one file. Not corrected.
- Various things are commented out of drama texts by caret brackets, including verse, quotations, and even sometimes some archaic word or phrase that was too difficult for the transcriber (More common, however, as noted above, is the commenting out of archaic forms and replacement in the text by [supposed] modern equivalents.) Not corrected.
- Where numeral *1* appeared as lower-case *L* <l> in dates, this has been corrected. Some attempt has been made to restore the pound character £ instead of *L* or *l*.. Many drama texts were scanned poorly; we have corrected the worst files using common sense, sometimes supplemented by online editions, though some garbled text resisted our efforts – see list below. We hope our rough-and-ready cleaning-up is an improvement for the time being, but it is not intended to be a substitute for proper proof-reading. Corrections should be sent to Heidelberg for collation in the next release.
- When a file was being edited anyway, we have often removed random indentation and superfluous inter-word spacing as well, and occasionally (in drama texts) filled in speaker gender and ensured that new speakers start on a new line. Many other files remain rather ragged, however.
- Please use the current filenames when citing or referring to ARCHER material.

Some individual file changes (not a complete list)

(Key for new texts typed in/collected by: HD = Heidelberg; MK = Manfred Krug's team; CM = Christian Mair's team; BK = Bernd Kortmann's team; RB = Richard Bailey's team.) For a complete list of files see **file_list_3-1.doc**, one long table showing all files in this release, very easily sorted or searched by year, author, genre, period and/or variety, or by old filename, and showing correspondences between new and old filenames.

new filename (old if removed)			change(s)
173x	fret	j3b	name change from 1735FRET.J2
1666	cav2	f2b	name change from 1666NEWC.F1A; lengthened (HD)
1670	durs	m2b	new text (CM)
1674	ano1	s2b	file was missing in ARCHER-1a
1674	gard	m2b	new text (CM)
1674	samp	m2b	new text (CM)
1675	hugy	s1	removed: translation
1675	leib	s1	removed: translation
1676	coxe	s2b	new text (CM)
1676	newt	s2b	new text (CM)
1678	morr	m2b	new text (CM)
1682	pr02	n1	removed: duplicated 1682pro2.n2b
1683	list	m2b	new text (CM)
1683	tyso	m2b	new text (CM)
1684	brig	m2b	new text (CM)
1684	wgmb	m2b	new text (CM)
1688	musg	m2b	new text (CM)
1692	cong	fc2b	lengthened (HD)
1698	sibb	m2b	new text (CM)
1699	dars	m2b	new text (CM)
1701	trot	d3b	lengthened (HD)
1710	pope	x2	removed
1720	dfoe	f3b	name change from 1720DEFO.F2
1728	rowe	f2	removed
1730	fiel	d3b	new text (HD)
1730	vanb	d3b	lengthened (HD)
1735	barr	m3b	new text (MK)
1735	gool	m3b	new text (MK)
1735	sim1	m3b	new text (MK)
1735	sim2	m3b	new text (MK)
1743	fiel	f3b	new text (HD)
1744	fldg	f3b	name change from 1744FIEL.F2
1752	lon2	n4b	<`> characters removed from text
1753	smol	fc3	removed
1762	publ	n4b	shortened approx. 800 words
1766	roge	d4a	new text (MK)
1769	bard	m4a	new text (BK)
1769	bart	s4a	new text (RB)

new filename (old if removed)			change(s)
1769	glos	m4a	new text (BK)
1769	norm	m4a	new text (BK)
1769	rush	m4a	new text (BK)
1769	west	s4a	new text (RB)
1770	munf	d4a	text at http://docsouth.unc.edu/southlit/munford/munford.html
1770	will	s4a	new text (RB)
1773	chal	m4b	new text (CM)
1773	perc	m4b	new text (CM)
1773	warr	d4a	new text (MK)
1774	hill	m4b	new text (CM)
1774	kell	m4b	new text (CM)
1775	ande	m4b	new text (CM)
1775	bath	m4b	new text (CM)
1775	fynn	m4b	new text (CM)
1775	mood	m4b	new text (BK)
1775	smit	m4b	new text (BK)
1775	whit	m4b	new text (BK)
1777	sher	d4b	new text (HD)
1785	fran	s4a	new text (RB)
1786	hopk	s4a	new text (RB)
1786	morg	s4a	new text (RB)
1786	perk	s4a	new text (RB)
1786	ritt	s4a	new text (RB)
1786	rus1	m4a	new text (BK)
1786	rus2	m4a	new text (BK)
1786	rush	s4a	new text (RB)
1786	wrig	m4a	new text (BK)
1789	low-	d4a	name change from 1789LOWE.D4, *Representative Plays by American Dramitists* corrected to *Dramatists*, corrections by common sense
1791	rush	s4a	new text (RB)
1793	hitc	f4a	revised (HD)
1793	smit	m4a	new text (BK)
1793	sta1	n4b	shortened approx 900 words
1793	sta2	n4b	shortened approx 1100 words
1794	rows	d4a	new text (MK)
1795	murd	d4a	new text (MK)
1796	sarg	d4a	new text (MK)
1799	deve	m4a	new text (HD)
1803	blak	x5	removed
1809	dimo	d5b	corrections by common sense
1813	poco	d5b	common sense, then last errors filled in from Lit Online
1815	aust	x5b	*tête-à-tête* restored
1819	moor	j5b	*January 1st, 1919* changed to *1819*
1820	aber	m5b	shortened approx 5000 words
1820	serl	d5b	a few corrections by common sense

new filename (old if removed)			change(s)
1825	barl	s5b	in some copies file was empty
1835	kenn	f5	removed: American
1836	marr	f5b	new text (HD); deleted hyphen in *closed-to*, added warning bracket, replaced angled double quotation marks by "
1839	plan	d5b	a few corrections by common sense, one stage direction still garbled
1844	bouc	d5b	accents restored; Lit Online Inconsistent and wrong glossing of <nuawt, nowt> by *not* removed
1845	surt	fc5	removed
1847	gask	f5b	shortened approx 3000 words
1849	arnd	x5b	£ restored and spacing tidied
1850	mlvl	f6a	name change from 1850MELV.F7
1851	dadd	s6a	new text (RB)
1857	hwth	x6a	stray <@> deleted
1861	elio	f6b	Lit Online; corrections by common sense
1863	tayl	d6b	shortened approx 1000 words
1864	bonn	m6b	shortened approx 3200 words
1864	mack	m6b	shortened approx 1400 words
1864	wats	m6b	shortened approx 5000 words
1867	robe	d6b	(Lit Online seems different); corrections by common sense
1869	hwls	x6a	just two corrections
1871	burr	s6a	new text (RB)
1871	lewi	d6b	http://gaslightmtroyalabca/thebellshtm – corrected and partially filled in
1873	elio	x6b	name change from 1873ELOT.X6
1877	jas-	f6a	name change from 1877JAME.F7
1878	hill	s6a	new text (RB)
1886	greg	s6a	new text (RB)
1886	whit	s6a	new text (RB)
1887	pres	s6a	new text (RB)
1889	madd	d6b	Lit Online; corrections by common sense
1891	holl	s6a	new text (RB)
1893	pine	d6b	shortened approx 1600 words; Lit Online; corrections by common sense
1894	holb	s6a	new text (RB)
1894	jone	d6b	shortened approx 700 words
1895	keel	s6a	new text (RB)
1895	shaw	d6b	Lit Online; corrections by common sense
1895	wild	d6b	Lit Online; corrections by common sense
1897	shaw	x6b	*Théatre Français* restored
1897	some	f6a	removed
1897	stur	s6a	new text (RB)
1899	mart	d6b	Lit Online; corrections by common sense
1908	jons	d7b	name change from 1908JONS.D8, corrections by common sense, two stage directions guessed
1908	yeat	d7b	corrections by common sense
1911	besi	d7b	corrections by common sense

new filename (old if removed)			change(s)
1917	firb	f7b	shortened approx 2500 words
1920	firb	d7b	corrections by common sense, quite a few stage directions still garbled
1922	fagn	d7b	corrections by common sense
1927	brow	f8a	removed
1935	brid	d7b	corrections by common sense, one stage direction still garbled
1935	gowr	d7b	corrections by common sense
1938	mccr	d7b	corrections by common sense, two stage directions still garbled
1943	haml	d7b	corrections by common sense, quite a few stage directions still garbled
1944	bagn	d7b	corrections by common sense, some stage directions still garbled
1951	fkn2	x8a	name change from 19512FKN.X0
1951	macl	x8a	*Arhibald* changed to *Archibald*
1953	ocon	x8a	one correction guessed, spacing tidied
1954	weav	s8a	new text (RB)
1955	hunt	s8a	new text (RB)
1955	ocsy	d8b	corrections by common sense, two stage directions still garbled
1956	mons	f9	removed
1958	lamm	f9	removed
1959	gua1	n8b	shortened approx 500 words
1960	bolt	d8b	http://wwwcooperedu/humanities/classes/coreclasses/hss2/library/man_for_all_seasonshtml; corrections by common sense
1960	mons	f9	removed: duplicated 1960cowa.f8b
1960	ratt	d8b	corrections by common sense
1961	gree	d8b	corrections by common sense
1965	macl	x8a	*Arhibald* changed to *Archibald*
1965	muel	s8a	new text (RB)
1965	new1	n8a	some missing words?
1966	jell	d8b	corrections by common sense
1966	mead	x0	removed
1967	deac	s8a	new text (RB)
1967	stm1	n8b	shortened approx 600 words
1969	bond	d8b	corrections by common sense
1969	ortn	d8b	corrections by common sense
1970	zind	d8a	filename corrected from either 1970ZIND.D0 or 1970FRNK.D0
1973	sinc	s8a	new text (RB)
1975	atl	n8a	some missing words?
1975	bish	s8b	oddities esp in connection with measurements
1976	broo	s8a	new text (RB)
1980	macl	x8a	*Arhibald* changed to *Archibald*
1982	chi1	n8a	shortened approx 1100 words
1982	chi2	n8a	shortened approx 800 words
1985	cree	m8b	new text (MK)
1985	evan	m8b	new text (MK)
1985	mack	m8b	new text (MK)
1987	magn	s8a	new text (RB)

new filename (old if removed)			change(s)
1988	smit	s8a	new text (RB)
1989	lat2	n8a	shortened approx 1000 words
1994	cart	s8a	new text (RB)
1997	krin	s8a	new text (RB)

The files 1713SARA.X2, 1714SARA.X2, 1715SARA.X2, 1722SARA.X2, 1739WWAY.X2, 1740WWAY.X2, 1743LAET.X2, 1749SFLD.X2, 1754SFLD.X2, 1773MRCY.X4, 1774LSMT.X4, 1775EUPN.X4, 1776EUPN.X4 (old names) were also removed, but these were included only in "ARCHER-1b"

David Denison, Sebastian Hoffmann and Nadja Nesselhauf