1. The Corpus

- Managed as an ongoing project by a consortium of participants at 14 universities in 7 countries. Since December 2008 it has been co-ordinated from Manchester.
- Reference line and consortium universities:

2. Design

Varieties
- British (b), all periods
- American (a), from 1750

Periods
- Divided into 50 year periods:

Genres
- advertising (a), drama (d), fiction (f), sermons (h), journal (j), legal (l), medicine (m), news (n), early prose (p), science (s), letters (x), diary (y) [NB. journal/diary (j) in previous versions]

Target sampling
- 10 texts, c. 2,000w each, per genre and variety in each period

3. Documentation

- Number of files and words per period, genre and variety
- Complete file list, with mapping to/from filenames in previous versions
- Complete word list, with frequencies
- Perl script for counting ‘words’
- List of non-ASCII characters and how they are coded
- Style sheet for XML reader
- Bibliographic database
- Website

Annotations
- All headers contain: (i) current filename; (ii) word count; (iii) bibliographic information

4. Versions of ARCHER

ARCHER-1
- Compiled 1990-93 by Douglas Biber and Edward Finegan
- Output: 10 different genres; British 1650-1990s; American 1750/1850/1950
- 3 slightly different versions: ARCHER-1 (Biber & Finegan), ARCHER-1a (German universities 2005), ARCHER-1b (Manchester 2005)

ARCHER-2
- Compiled in the early 2000s, as based on ARCHER-1; completed in 2004-05
- Expanded by filling gaps in the American variety and adding some British files for 1600-49
- Output: ARCHER-1 plus new texts; one new genre (Advertising, American only); one more period (1600-49) for some genres

ARCHER-3.1
- Compiled in 2006, co-ordinated from Heidelberg
- Aimed to obtain a more balanced corpus by (i) temporarily excluding genres that did not have a BrE or AmE counterpart; (ii) eliminating inconsistencies in the previous versions; (iii) adding new texts
- Output: ARCHER-1b revised; some new texts; some materials from ARCHER-2 removed

ARCHER-3.2
- Compiled in 2013, co-ordinated from Manchester since 2008
- Expansion of periods and genres by (i) restoring files omitted from ARCHER-3.1 but included in ARCHER-1 and ARCHER-2; (ii) splitting the single category journals-diaries into two: journals (j) vs. diaries (y); (iii) adding new texts
- Textual accuracy and consistency in the provision of bibliographic information has also been improved
- Output: ARCHER-3.1, plus ARCHER-1 and ARCHER-2 files excluded from version 3.1, plus new texts
- One version morpho-syntactically tagged with CLAWS5
- An additional version tagged, chunked and parsed with Treebank conventions

Table 1. Versions of ARCHER

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>version 1b</td>
<td>29</td>
<td>56</td>
<td>630</td>
<td>2,675</td>
</tr>
<tr>
<td>words</td>
<td>1.3 million</td>
<td>1.3 million</td>
<td>1.3 million</td>
<td>1.3 million</td>
</tr>
<tr>
<td>genres</td>
<td>(a, d, f, h, j, l, m, n, s, x)</td>
<td>(d, f, h, j, l, m, n, p, s, x)</td>
<td>(d, f, h, j, l, m, n, p, x)</td>
<td>(d, f, h, j, l, m, n, p, x)</td>
</tr>
<tr>
<td>British English</td>
<td>31</td>
<td>31</td>
<td>13,500</td>
<td>1.3 million</td>
</tr>
<tr>
<td>words</td>
<td>160,000</td>
<td>160,000</td>
<td>1650–1990</td>
<td>1650–1990</td>
</tr>
<tr>
<td>genres</td>
<td>(a, d, f, h, j, l, m, n, s, x)</td>
<td>(d, f, h, j, l, m, n, p, s, x)</td>
<td>(d, f, h, j, l, m, n, p, x)</td>
<td>(d, f, h, j, l, m, n, p, x)</td>
</tr>
<tr>
<td>American English</td>
<td>12</td>
<td>155</td>
<td>31</td>
<td>31</td>
</tr>
<tr>
<td>words</td>
<td>1.9 million</td>
<td>3,524,000</td>
<td>1.8 million</td>
<td>1.8 million</td>
</tr>
<tr>
<td>genres</td>
<td>(a, d, f, h, j, l, m, n, s, x)</td>
<td>(d, f, h, j, l, m, n, p, s, x)</td>
<td>(d, f, h, j, l, m, n, p, x)</td>
<td>(d, f, h, j, l, m, n, p, x)</td>
</tr>
</tbody>
</table>

ARCHER format
- Untagged plain text
- Non-POS-tagged XML
- POS-tagged XML version
- POS-tagged files prepared and indexed for CQPweb

Figure 1. ARCHER-3.1 (2006)

British English (674 files – c. 1,253,000 million words)

Figure 2. ARCHER-3.2 (vsn3.1 + new material, March 2013)

British English (1,075 files – c. 1,958,000 million words)

Further information: