ARCHER 3.2

A Representative Corpus of Historical English Registers

Nuria Yáñez-Bouza
www.manchester.ac.uk/archer

1. The Corpus

- A multi-genre historical corpus of written and speech-based British and American English, 1600-1999.
- 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.
- Versions: ARCHER-1 (1992-93), ARCHER-2 (2004-05), ARCHER-3.1 (2006), ARCHER-3.2 (2013).
- Reference line and consortium universities:

  ARCHER-X = *A Representative Corpus of Historical English Registers* version X. 1990–1993/2002/2007/2010/2013. Originally compiled under the supervision of Douglas Biber and Edward Finegan at Northern Arizona University and University of Southern California; modified and expanded by subsequent members of a consortium of universities. Current member universities are Northern Arizona, Southern California, Freiburg, Heidelberg, Helsinki, Uppsala, Michigan, Manchester, Lancaster, Bamberg, Zurich, Trier, Santiago de Compostela and Leicester.

2. Design

Varieties

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

Periods

- Divided into 50 year periods
- 1600–49 (1), 1650–99 (2), 1700–49 (3), 1750–99 (4), 1800–49 (5), 1850–99 (6), 1900–49 (7), 1950–99 (8)

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, genreand 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
- TEI-headers: file, encoding and profile description plus revision history
- morpho-syntactically tagged with CLAWS7
- consistent mark-up for speakers in fiction and characters in drama

4. Versions of ARCHER

ARCHER-1

- compiled 1990-93 by Douglas Biber & 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

- completed 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

- completed 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 CLAWS7
- an additional version tagged, chunked and parsed with Treebank conventions

Table 1. Versions of ARCHER

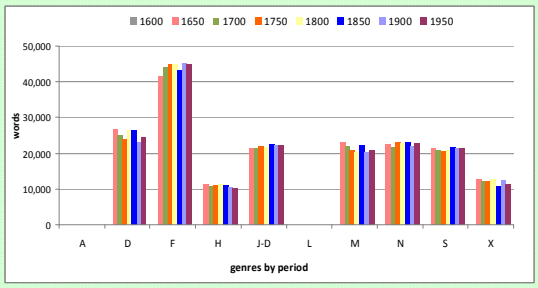
	ARCHER 1 (1992-93) Version 1b (Manchester)	ARCHER 2 (2004-05), new files	ARCHER 3.1 (2006)	ARCHER 3.2 (2013)
BRITISH ENGLISH				
files	664	20	674	1,075
words	c. 1.3 million	c. 63,000	c. 1.3 million	c. 2 million
periods	1650-1990	1600-49	1650-1999	1650-1999 (1600-1649 also in d, l, p)
genres	8 (d, f, h, j, m, n, s, x)	2 (d, p)	8 (d, f, h, j, m, n, s, x)	12 (a, d, f, h, j, l, m, n, p, s, x, y)
AMERICAN ENGLISH				
files	298	92	281	635
words	c. 60,000	c. 331,000	c. 535,000	c. 1.3 million
periods	1750-99, 1850-99, 1950-90 Legal 1750-1990	1750-99, 1850-99, 1950-90 Adv 1750-1990	1750-99, 1850-99, 1950-99	1750-1999
genres	8 (d, f, h, j, l, m, n, x)	4 (a, d, f, n)	8 (d, f, h, j, m, n, s, x)	11 (a, d, f, h, j, l, m, n, s, x, y)
TOTAL				
files	962	112	955	1,710
words	c. 1.9 million	c. 394,000	c. 1.8 million	c. 3.3 million
periods	1650-1990	1600-1990	1650-1999	1650-1999 (1600-1649 also in d, l, p)
genres	9 (d, f, h, j, l, m, n, s, x)	5 (a, d, f, n, p)	8 (d, f, h, j, m, n, s, x)	12 (a, d, f, h, j, l, m, n, p, s, x, y)

ARCHER format

- untagged plain text (extension .txt)
- non-POS-tagged XML (extension .xml)
- POS-tagged XML version (extension .c7x)
- POS-tagged files prepared and indexed for CQPweb

Figure 1. ARCHER-3.1 (2006)

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



American English (281 files – c. 536,000 words)

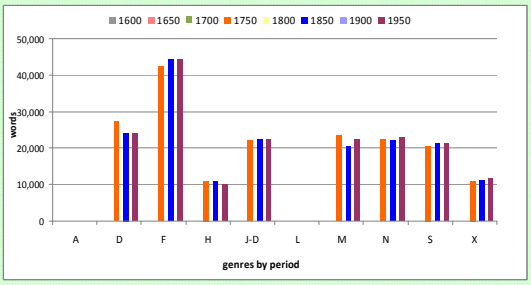
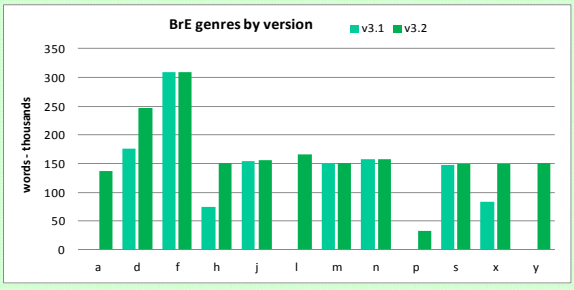
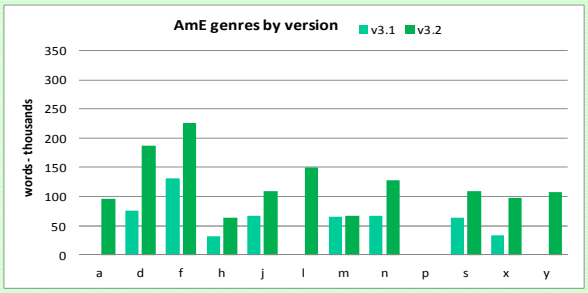


Figure 2. ARCHER-3.2 (vsn 3.1 + new material, March 2013)

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



American English (635 files – c. 1,340,000 million words)



Further information:
Biber, D., E. Finegan & D. Atkinson. 1994. ARCHER and its challenges: Compiling and exploring a representative corpus of historical English registers. In U. Fries, G. Tottie & P. Schneider (eds.). *Creating and using English language corpora*. Amsterdam: Rodopi, 1-13.
Yáñez-Bouza, Nuria. 2011. ARCHER past and present (1990-2010). *ICAME Journal* 35, 205-236

Acknowledgements:
David Denison, Sebastian Hoffmann and Nadja Nesselhauf.
British Academy Small Research Grant (SG-101087), The University of Manchester, 2010-12.