ARCHER past and present (1990-2011)

A Representative Corpus of Historical English Registers

Nuria Yáñez-Bouza

School of Languages, Linguistics and Cultures, The University of Manchester

1. The Corpus

- A multi-genre historical corpus of written and speech-based British and American English, 1600-1999.
- In in-house use and 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), the current version**, and ARCHER-3.2 (in progress).
- Access on-site at the consortium universities:

  ARCHER-3.1 = A Representative Corpus of Historical English Registers 3.1. 1990 1993/2002/2007/2010. Compiled under the supervision of Douglas Biber and Edward Finegan at Northern Arizona, Southern California, Freiburg, Heidelberg, Helsinki, Uppsala, Michigan, Manchester, Lancaster, Bamberg, Zurich, Trier, Salford, and Santiago de Compostela.

2. Design

Varieties

- British (b), all periods
- American (a), often the latter part of the century only

Periods

- 1600-1999, divided into 50 year periods
- pre-1600 (0), 1600-49 (1), 1650-99 (2), 1700-49 (3), 1750-99 (4), 1800-49 (5), 1850-99 (6), 1900-49 (7), 1950-99 (8), post-2000 (9)

Genres

- advertising (a), drama (d), fiction (f), sermons (h), journal/diary (j) [OR journal (j), diary (y)], legal (l), medicine (m), news (n), science (s), letters (x)

Target sampling

- 10 texts, c. 2,000 words each, per genre and variety in each period

3. Documentation

- tables with the 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
- bibliographic database
- website

Annotations:

- all headers contain: (i) the current filename; (ii) word count; (iii) bibliographic information
- bibliographic annotations in the sample headers and in text
- no systematic coding for sociolinguistic information
- neither tagged nor parsed

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 in circulation: ARCHER-1 (Biber & Finegan), ARCHER-1a (German universities 2005), ARCHER-1b (Manchester 2005)

ARCHER-2

- compiled in the early 2000s; 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 2004-06, co-ordinated from Heidelberg (2004-08)
- 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, plus new texts, minus ARCHER-2

ARCHER-3.2

- under compilation (2008-), co-ordinated from Manchester (2008-)
- correction and expansion of periods and genres by (i) restoring files omitted from ARCHER-3.1; (ii) splitting the single category journals-diaries into two: journals (j) vs. diaries (y); (iii) adding new texts
- morpho-syntactic tagging with CLAWS8
- output: ARCHER-3.1, plus ARCHER-2 and ARCHER-1 files excluded from vsn 3.1, plus new texts

Table 1. Versions of ARCHER

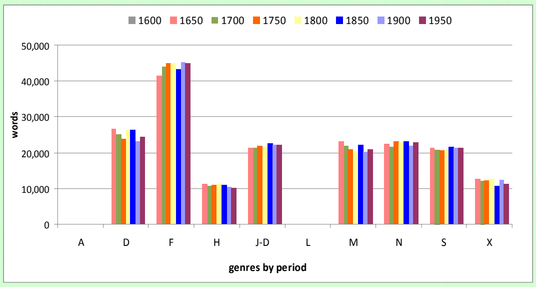
		ARCHER-1 (1992-93) version 1b	ARCHER-2 (2004-05) new data only	ARCHER-3.1 (2006)	ARCHER-3.2 (forthc.)
BrE	files	664	20	674	c. 1060
	words	c. 1.3 million	c. 64,000	c. 1.3 million	c. 1.9 million
	period	1650-1990	1600-49	1650-1999	1600-1999 (only d, f, l 1600-49)
	genres	8 (d, f, h, j, m, n, s, x)	2 (d, f)	8 (d, f, h, j, m, n, s, x)	11 (a, d, f, h, j, l, m, n, s, x, y)
AmE	files	298	92	281	c. 580
	words	c. 60,000	c. 330,000	c. 500,000	c. 1.3 million
	period	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	c. 1,640
	words	c. 1.9 million	c. 390,000	c. 1.8 million	c. 3.2 million
	period	1650-1990	1600-1990	1650-1999	1600-1999
	genres	9 (d, f, h, j, l, m, n, s, 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)

ARCHER-3.1 format

- basic version: 955 separate text files, in the Windows Latin 1 character set
- 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
- a version with only 7-bit ASCII characters used, containing HTML-like entities instead of single characters, in a zip archive

Figure 1. ARCHER-3.1 (2006)

1.a. British English (674 files – c. 1.3 million words)



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

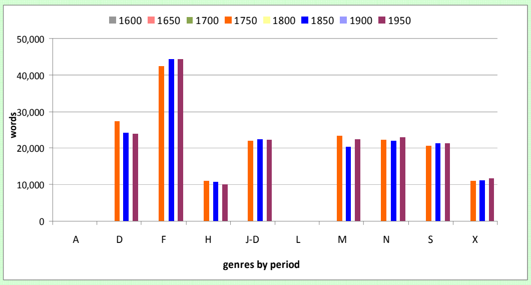
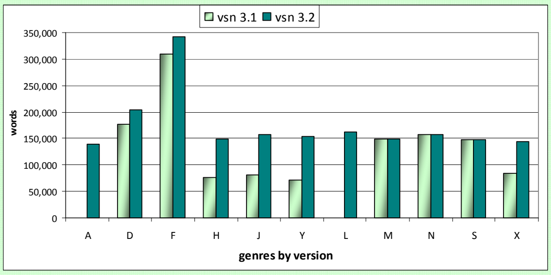
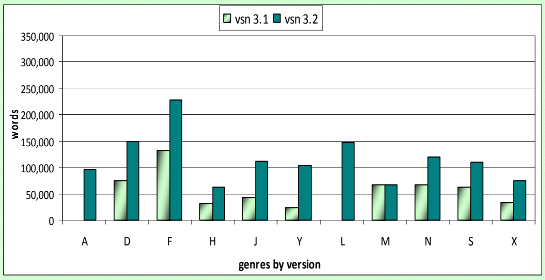


Figure 2. ARCHER-3.2 (vsn 3.1 + new material, April 2011)

2.a. British English (c. 1,060 files – c. 1.9 million words)



2.b. American English (c. 580 files – c. 1.3 million words)



Further information:

ARCHER web page: www.llc.manchester.ac.uk/research/projects/archer/

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.

Acknowledgements:

David Denison, Sebastian Hoffmann and Nadja Nesselhauf.

British Academy Small Research Grant (SG-101087), The University of Manchester, 2010-12.

nuria.yanez-bouza@manchester.ac.uk