ARCHER past and present (1990–2010)[1]

Nuria Yáñez-Bouza, The University of Manchester

1 Introduction

ARCHER (*A Representative Corpus of Historical English Registers*) is a multi-genre historical corpus of ca. 1.8 million words of British and American English covering the period 1650–1999. First constructed by Douglas Biber and Edward Finegan in the early 1990s at the universities of Northern Arizona and Southern California, it is now in in-house use and managed as an ongoing project by a consortium of participants at fourteen universities in seven countries (see Section 4.3). Since December 2008 it is being co-ordinated from Manchester (UK). Although the corpus has been used for a large number of studies in the history of English, and although it already exists in three versions (ARCHER-1, ARCHER-2, ARCHER 3.1) and is on its way towards the fourth version (ARCHER 3.2), a comprehensive account of its structure and contents since its early days is lacking (cf. Biber *et al.* 1994a, 1994b; Biber and Finegan 1997: 255–257). This paper will fill the gap in the literature by telling the history of ARCHER from 1990 to 2010.

The approach is two-fold: past and present. Section 2 offers an overview of the background of the corpus, its aims and design. In Section 3, I describe the contents and structure of ARCHER-1 and the additions in ARCHER-2. Section 4 focuses on the current version, ARCHER 3.1, including (i) an account of the changes in the structure and contents of this version as compared to earlier versions (Section 4.1); (ii) a summary of the new coding conventions for text annotation and of the edits carried out, primarily with regard to filenames (Section 4.2); and (iii) remarks on access to the corpus (Section 4.3). The paper will close with an outlook of the ongoing phase towards ARCHER 3.2: aims, work in progress and expectations (Section 5).

2 Aims and design

ARCHER was first constructed in the early 1990s in the heyday of compilation of diachronic corpora for the study of the English language, largely influenced by the successful completion of the *Helsinki Corpus* in the late 1980s. See, for instance, the *Corpus of Late Modern English Prose*, the *Zurich English Newspaper Corpus* (ZEN), ICAMET, the *Century of Prose Corpus*, the *Corpus of Irish English*, and others reported in Kytö, Rissanen and Wright (1994, esp. p. vii), Kytö and Rissanen (1995), and later Hickey *et al.* (1997). Their shared goal was to make available the (re)sources that would enable research primarily from a variationist perspective: language change as conditioned by genre, register, dialect or social class, as well as individual authors' styles.

In a similar way to the *Helsinki Corpus*, ARCHER was designed as a multipurpose diachronic corpus representing a wide range of register diversity across continuous historical periods (Biber *et al.* 1998: 251–253). With a diachronic coverage from 1650 to the 1990s, it would nicely fill the gap between the *Helsinki Corpus* (850–1710) and the various extant 20th-century corpora (e.g. BROWN, FROWN), thereby allowing matched comparison between British and American sub-corpora over a long time-frame. Besides, with a bit of manual reorganisation, text type continuity would also be achievable to a large extent, e.g. personal letters, sermons, legal texts or medicine.

More precisely, the background of ARCHER lies in Biber and Finegan's work on diachronic relations among oral and literate registers based on the framework of multi-dimensional analysis, whereby distributional patterns among linguistic features are investigated (see e.g. Biber 1988, Biber and Finegan 1988, 1989, 1992; and after the first version of the corpus had been completed, Biber 1995, Biber and Finegan 1997, Biber 2001). As they explained, ARCHER

> was designed for a specific major research agenda: to analyze historical change in the range of written and speech-based registers of English from 1650 to the present. The general design goal for the corpus has thus been to represent a wide range of register variation, sampled systematically across texts from the last three and a half centuries. (Biber and Finegan 1997: 255)

Thus, at the outset of the project, ARCHER was intended to facilitate research in the following areas:

(i) the historical evolution of written and speech-based registers in English, considered individually and relative to the overall patterns of change;

(ii) diachronic descriptions of the linguistic characteristics of speech-based and written registers of English across more than three centuries of Modern English;

(iii) synchronic descriptions of the linguistic characteristics of speech-based and written registers for each of several periods of Modern English;
(iv) comparative developments in British and American English;
(v) functional analyses of individual linguistic features;
(vi) co-occurrence patterns in lexicon and grammar;
(vii) changes in the social patterns of language use (especially gender variation) over time.

(adapted from Biber *et al.* 1994a: 2–3, Biber *et al.* 1994b: 4)

The corpus has been widely used in historical linguistics of the Modern English period for morphological, lexical, syntactic and stylistic research of all kinds; for some illustrative examples of studies carried out by the compilers in the early days, see Biber *et al.* (1998: 203–227).

Text type categorisation is crucial in designing a multi-genre diachronic corpus like ARCHER. Three main criteria were borne in mind: (i) the selection of registers representative of a wide range of writings in a given historical period; (ii) the selection of registers that had a continuous history across periods; and (iii) the inclusion of speech-based registers that, within their known limitations, would reflect the characteristics of spoken language in the historical period under investigation (see Biber *et al.* 1998: 252). The written registers include personal styles of communication (journal/diaries and personal letters), fiction prose, popular exposition represented by news reportage, and specialist expository registers illustrated with legal opinions, medical prose, and scientific prose. Speech-based registers are represented by sermons as a reflection of planned monologue styles, and dialogue in drama and fiction as reflections of casual face-to-face conversation (see Biber and Finegan 1997: 255–257).[2] ARCHER also aims to represent both formal and informal kinds of writing (Biber *et al.* 1994a: 3). At the informal end are journal/diaries and letters among the written registers, and drama and fiction dialogue among the speech-based registers. The more formal end is represented by legal opinions, medicine and science, on the one hand, and sermons, on the other. Fiction prose and news stand between the two poles in the continuum. (See Section 3 on changes across versions.)

British registers are sampled in seven 50-year periods from 1650 to 1990/99, while American English registers were to be sampled, in the first instance, for only three, the second halves of the 18th, 19th and 20th centuries. As explained by Biber and Finegan, “the lesser sampling of American texts was motivated not by theoretical considerations but by expedience, in response to a task that turned out to be bigger than the available resources” (1997: 273, note 2).

In most cases the texts were chosen using random sampling techniques from available bibliographies (see further Biber *et al.* 1994a: 4–7). The selected texts were written in prose, with very few exceptional cases of verse dialogue in drama and fiction, and they were scanned or typed in from editions rather than manuscripts. Generally, texts consist of continuous passages from the same source. However, at times the compilers opted for sampling passages from the beginning, middle and end, or for including two/three texts from different sources (e.g. British science). The target sampling is ten texts of at least 2,000 words per genre and variety in each 50-year period. However, as is well-known to the corpus linguist, symmetry is not always attainable, especially because of the varied number and quality of available texts in some periods and/or genres. In some genres we have to do with short texts, particularly personal letters, or smaller samples altogether, like sermons; on the other hand, fiction texts tend to be bigger in order to include a representative number words of both narrative and dialogue passages. Typically, a full sampling for a genre would include 100 texts, 20,000 words. As Finegan and Biber (1995: 244) acknowledged, the sample size may look small by current standards of synchronic corpora, but other small corpora like the *Helsinki Corpus* (ca. 1.6 million words) have undoubtedly contributed a great deal to the knowledge of the history of the English language. As the compilers of the *Helsinki Corpus* put it, its diachronic and textual coverage is "extensive enough to show fairly reliable and consistent trends of development in a large number of topics" (Kytö and Rissanen 1993: 4); admittedly, more so for morphological and syntactic analyses than for lexical investigations, but even in the latter studies tentative trends can be derived, too (see, to name a few, Rissanen *et al.* 1993, Rissanen *et al.* 1997). Users, they continue, "can easily sharpen the picture given by our corpus with details obtainable from the texts themselves, from other corpora [or] from printed concordances" (Kytö and Rissanen 1993: 4). We like to think of ARCHER in the same way: an adequately representative corpus, as its name suggests, large enough to give statistically significant results even at microscopic level (see Biber *et al.* 1998: 203–227).[3]

Table 1 offers an overview of the target design of ARCHER as adapted from Biber and Finegan (1997: 256), a revised version of Biber *et al.* (1994b: 5). Section 3 will describe the development of the corpus to the present day.

Table 1: Overall design of ARCHER at the outset (adapted from Biber and Finegan 1997: 256)

Time-span:	1650–1990, divided into 50 year periods
Varieties:	British (all periods) and American (one period per century)
Genres/Registers:	seven written categories: journals/diaries, personal letters, fiction prose, news reportage, legal opinions, medical prose, scientific prose three speech-based categories: drama, fiction dialogue, sermons
Target sampling:	ten texts, at least 2,000 words, per genre and variety in each period A full sampling for a genre would include 100 texts:

1650–99	British:	10 texts		
1700–49	British:	10 texts		
1750–99	British:	10 texts	American:	10 texts
1800–49	British:	10 texts		
1850–99	British:	10 texts	American:	10 texts
1900–49	British:	10 texts		
1950–90	British:	10 texts	American:	10 texts

3 *Past: ARCHER-1 and ARCHER-2*

Over its twenty-year life span, ARCHER exists in three versions and is on its way towards the fourth. The earliest version, referred to as ARCHER-1, was compiled in 1990–1993 by Douglas Biber (Northern Arizona University) and Edward Finegan (University of Southern California). New texts were compiled in the early 2000s primarily with a view to filling the gaps in the coverage of the American variety; ARCHER-2 was completed in 2004–2005. The third phase of the project began in 2004 under the co-ordination of Heidelberg (2004–2008). The aim was to obtain a more balanced corpus in terms of genre distribution by (i) (temporarily) excluding genres that did not have a British or American counterpart; (ii) adding new texts; and (ii) eliminating inconsistencies in the previous versions (e.g. duplicate texts); ARCHER 3.1 was completed in summer 2006. The phase 2008–2010 is a new stage of expansion, with the added value of developing a tagged version of the corpus; the outcome will be ARCHER 3.2.

As reported by Biber and Finegan (1997: 255–257), amongst others, the overall structure of the earliest version of ARCHER contained ten major register categories, grouped in 50-year periods from 1650 to 1990, and altogether the corpus consisted of 1,037 texts and ca. 1.7 million words. Table 2 displays the breakdown of texts by genre/register, including the total number of sample texts compiled, input and tagged:

Table 2: ARCHER-1: Breakdown of texts by genre/register (adapted from Biber and Finegan 1997: 257)

Genre/Register	**Remarks**	**Number of texts collected, input and tagged**
Drama	only 5 texts from 18th-century American	95
Fiction prose		100
Fiction dialogue		100
Journals/Diaries		100
Legal opinions	1750–1990 but American only	57
Letters	more than 10 texts per period; most texts shorter than 1,000 words	275
Medicine	no 18th-century American samples	90
News		100
Science	British only; from *Philosophical Transactions of the Royal Society*	70
Sermons	only 5 texts per period	50

However, detective work resulting from digging into past documentation for this paper has brought to light the existence of three slightly different versions of ARCHER-1, namely ARCHER-1 as just described and reported in other studies (e.g. Biber 2001: 94–95, 2004: 197–199); ARCHER-1a, hosted at German universities mainly, where many files, especially fiction, had apparently been thoroughly corrected at some point; and ARCHER-1b, the version held at Manchester in 2005, when the corpus was first documented in full (see Section 5.3). In comparison with ARCHER-1, ARCHER-1b had more words but fewer texts: ca. 1.9 million from 962 samples altogether. Other differences observed include: (i) the compilation of fiction as a single genre rather than divided into 'fiction prose' and 'fiction dialogue'; (ii) 11 extra fiction files; (iii) 13 extra letter samples; and (iv) one extra sample from news. On the other hand, it must be mentioned that all the remarks made in Table 2 hold true for ARCHER-1b, too. The differences between ARCHER-1a and ARCHER-1b were already uncovered during the preparation of ARCHER 3.1 in 2006. In short, 17 files were missing in the former (13 letters, 3 fiction texts, 1 science text), and the science text 1825barl.s5b was empty in ARCHER-1a but filled with text in ARCHER-1b. The two versions were tidied up and combined in ARCHER 3.1, except for the letters, which will be restored in ARCHER 3.2.[4]

It must be emphasised that the acknowledgement of these differences is not intended to diminish the value of the corpus or of the results obtained from it at different institutions. On the contrary, it is hoped that it will clarify/explain (potential) small discrepancies in the findings obtained by different scholars and/or at different points in the history of ARCHER. We will aim for ARCHER 3.2 to comprehend all (suitable) files extant in every early version of the corpus.

The second version of the corpus is known as ARCHER-2 (2004–05). The version at Manchester comprised ARCHER-1b in its entirety plus ca. 394,000 words from 92 new American texts and 20 British texts. The latter only involved early drama and fiction from 1600–49, a period which to date remains available for these genres and in this version alone. American texts from drama, fiction and news reportage were sampled for the periods 1800–49 and 1900–49, hence achieving continuity from 1750 to 1990. Another improvement was the incorporation of a new genre, advertising, also covering 1750 to 1990, albeit with American samples only and with very few texts in the early subperiods.[5] With advertising, the total number of registers increased to 11. Advertising falls into the category of written registers, half way along the oral/literate continuum. Altogether, in 2005 ARCHER-2 consisted of 1,074 files and ca. 2,300,000 words (see Table 3 in Section 4.1).

4 Present: ARCHER 3.1

4.1 A phase of revision

The purpose of ARCHER 3.1 was to enhance the usefulness of the corpus in a number of ways, from the revision of genre, period and geographical coverage to specific file correction. An added value to this version was the compilation of a comprehensive database with full bibliographic information for every file that belongs or has belonged to ARCHER.

ARCHER 3.1 consists of ca. 1.8 million words from a total of 955 files; that is over 1.2 million words from 674 British files and over half a million words from 281 American files (see Table 3 and Appendix). This version incorporates a partial clean-up and correction of many files and deletion of some texts from previous versions, as well as additions of new texts. All changes have been carefully documented in the ARCHER database, keeping track of correspondences between new and old files. The complete list of edits, which concern approximately 190 files, is not appendixed to this paper for reasons of space, but is available at the consortium universities and can also be obtained from the Manchester team on request. The main edits are summarised as follows:

(i) All new files in ARCHER-2 were excluded. As described above, these involve 20 British files (drama, fiction) and 92 American files (advertising, drama, fiction, news reportage), ca. 394,000 words in total.

(ii) American legal texts from ARCHER-1 were likewise excluded *in toto*: 57 files, ca. 147,000 words.

(iii) 28 individual files from ARCHER-1 and ARCHER-2 were removed for various reasons, such as duplicate files, samples missing in the documentation of some universities, or unbalanced word counts and number of texts in certain periods/genres. The files affected are one duplicate text from news reportage, two science files (both translations), nine fiction files, and 16 letters, 13 of which were already missing in version ARCHER-1a. One fiction file was removed because it had been misclassified as British rather than American (1835kenn.f5).

(iv) 17 samples were shortened in order to achieve a more balanced distribution per genre and/or period: two from British fiction, three from British drama, four from British medicine, five from British newspapers and three from American newspapers. The number of words excised ranges from 500 to 5,000.

(v) For similar reasons, four British files were lengthened: two from fiction and two from drama.

(vi) One fiction text in particular was fully revised (1793hitc.f4a), while some other 50 samples were tidied up in one way or another, e.g. inconsistent use of angle brackets, punctuation, mark-up, etc.

(vii) 78 new files were added to the remaining body of the corpus, adding up to over 157,000 words (ca. 99,000 words of American English, 44 files, and ca. 58,000 of British English, 34 files). The new files concern American science (30), British science (2), American medicine (9), British medicine (28), American drama (5), British drama (2), and British fiction (2). To these, we must add the incorporation of four British files which were missing in version ARCHER-1a: two from science and two from fiction.

As a result of these changes, ARCHER 3.1 presents itself as a more balanced corpus covering eight different genres across the two varieties of English from 1650 to 1999 (see Appendix). Following Biber and Finegan's original categorisation (Section 2), ARCHER 3.1 contains drama and sermons as speech-based text types, along with fiction dialogue extracts; personal letters and journals/diaries as written categories but relatively close to speech-based categories; and fiction prose, medicine, science and news reportage as written types of text. Note that American legal opinions (ARCHER-1) and American advertising (ARCHER-2) are being held over until the forthcoming ARCHER 3.2, both with new additions to the corresponding British English variety (see Section 5). Regarding chronological coverage, the British variety in ARCHER 3.1 ranges continuously from 1650 to 1999 in all genres, while the American variety is sampled in three subperiods only, 1750–99, 1850–99, 1950–99, but now for all the genres represented. Given the exclusion of the texts added to ARCHER-2, the periods 1800–49 and 1900–49 in American news, drama and fiction are no longer represented in this version. For the same reason, the early 17th-century period (1600–49) is not sampled either in the British variety. They will all be restored in the forthcoming ARCHER 3.2.

Table 3 summarises the coverage of the three complete phases as documented at Manchester in October 2010, giving details of files, words, genres and periods per variety. Notice that the column for ARCHER-2 includes new data only, as it builds upon ARCHER-1. The tables in the Appendix offer a fuller account of the corpus by displaying the number of files and words for every period and genre, tabulated separately for British and American varieties, with subtotals for each period and each genre.

Table 3: ARCHER versions 1992/93–2006: number of files and words per project phase[6]

		ARCHER-1 (1992–93) version 1b	**ARCHER-2 (2004–05) new data only**	**ARCHER 3.1 (2006)**
BrE	files	664	20	674
	words	1,299,670	63,274	1,253,557
	genres	8	2	8
	periods	1650–1990	1600–49	1650–1999
AmE	files	298	92	281
	words	594,041	330,822	535,752
	genres	8	4	8
	periods	1750–99, 1850–99, 1950–90 *Legal* 1750–1990	1750–99, 1850–99, 1950–90 *Advertising* 1750–1990	1750–99, 1850–99, 1950–99
Total	files	962	112	955
	words	1,893,711	394,096	1,789,309
	genres	9	4	8
	periods	1650–1990	1600–1990	1650–1999

The materials of ARCHER 3.1 became accessible to the consortium in 2006 in various convenient formats and with up-to-date documentation, including a complete list of files, new word counts, a new labelling system for filenames, and a suggested protocol for future compilation of texts. These issues are described in the following subsections.

4.2 Coding conventions

With a view to achieving greater consistency and coherence throughout, a set of guidelines for compilation and annotation of texts was drawn up and distributed with ARCHER 3.1. It must be noted that not every file in ARCHER 3.1 adheres to these conventions rigidly, as sufficient resources were not available to carry out an in-depth revision by the time the corpus was completed; this is work in progess towards the enhanced version ARCHER 3.2.

4.2.1 Annotations

Unlike other historical corpora, e.g. the *Helsinki Corpus* or the *Corpus of Early English Correspondence*, ARCHER has not been coded in COCOA format, nor does it contain sociolinguistic information. The author’s gender, and occasionally the addressee’s, has been documented but only in the database, available on request. ARCHER does contain bibliographic annotations in the sample headers and also in-text annotations made by the compiler(s). All material inserted by corpus compilers or editors appear within angle brackets, <>. Outside the headers, existing angle brackets have been retained for the time being as they were, however. In some texts they enclose original spellings as opposed to normalised spellings (see bold strings in Sample 1). (This practice is the logical opposite of what we would suggest for a historical corpus and is therefore a priority task for correction towards ARCHER 3.2.) In some drama texts, angle brackets also enclose original stage directions, verse, quotations, and even sometimes some archaic word or phrase that was too difficult for the transcriber. Unpaired brackets in earlier versions of the corpus have been corrected for angle <>, curly {} and square [], but not yet for round () brackets.

<1671cary.d2b, 3706 words, was 1671CARY.D1; 1671CARYLL.D1>
<Caryll, John. 1671. Sir Salomon; or The Cautious Coxcomb.>

{=m Sir Salo.} **Precious <Pretious>** Coxcombes; Open the Door
quickly, or **I'll <I'le>** make you both fast this **sevennight**
<se'night> from Beef and Pudding.

{=m Sir Salo.} Peace, both of you; Will you never arrive to
Common **sense <sence>**?

{=m WARY.} <two lines of verse>
<Begone, my reasons are but lost on thee:
For no dispute can cure Love's Heresie.> <Exit Julia.>

Sample 1: Use (and misuse) of angle brackets: extracts from a 17th-century drama text (1671cary.d2b)

4.2.2 Headers

All headers contain at least three pieces of information: (i) the current filename in the formula nnnnabcd.gpv; (ii) an accurate word count; and (iii) bibliographic information about the sample text and the edition from which it has been taken. If the file is not completely new but dates back from before ARCHER 3.1, the previous filename will be retained later in the header, too, sometimes with even older variants retained from earlier incarnations. When the file contains samples from more than one author, as often happens in science, we may find that each sample has its own header with individual bibliographic information and word count. Other annotations the sample header may include concern footnotes and/or the use of italics; for instance <Footnotes are not represented>, as in Sample 2, or <First paragraph is reverse indented and in italics>. Many headers were corrected when the texts were being verified for ARCHER 3.1, and the format has been partly regularised.

```
<1777sher.d4b, 2581 words, was 1777SHER.d3; 1777SHERIDAN.D3;
1777SHER.D3.>
<D3. Sheridan, Richard Brinsley.
The School for Scandal. A Comedy, in Five Acts. First Performed at the The-
atre Royal Drury Lane, on 8th May, 1777.
London: Thomas Hailes Lacy.>
<Footnotes are not represented.>
<>

<AU=F>
<CHARACTERS>
<=m Sir Peter Teazle>
<=m Crabtree>
<=m Sir Benjamin Backbite>
<=m Joseph Surface>
<=m Servant>
<=f Lady Teazle>
<=f Lady Sneerwell>
<=f Mrs. Candour>
<=f Maria>
<=m =f indicate gender>

         <ACT II.>
      <SCENE I. – Same as last, but in 1st grooves.>
   <Enter SIR PETER and LADY TEAZLE, following, R.>
```

Sample 2: Headers in ARCHER 3.1: sample from an 18th-century drama file (1777sher.d4b)

4.2.3 Filenames

One of the major tasks and innovations in ARCHER 3.1 was the revision and correction of filenames from the original formula DATENAME.RP to nnnnabcd.gpv. As explained by Biber *et al.* (1994a: 4), each filename in ARCHER-1 indicated date, author, register and period in the form DATENAME.RP, where DATE reflects the date of publication or (in the case of letters and diaries) composition; NAME is the author identification; R stands for register; and P represents the abbreviation for the relevant 50-year period. To use their example, the filename 1881BESA.F6 indicates that this is a text written in 1881 by the author Water Besant, is fiction (F), and belongs to period 6 (1850–99).

The main problem with the formula DATENAME.RP is that it is not transparent about which variety the sample belongs to, whether British or American, because P does not follow a chronological sequence but, rather, a combination of chronological and geographical differentiation. The first two periods were represented by British samples only (1= 1650–99, 2 = 1700–49). The extensions 3, 6 and 9 identified British texts for the periods 1750–99, 1850–99 and 1950–99, respectively, parallel to the American texts with the extensions 4, 7 and 0 (i.e. 10). The extensions 5 (1800–49) and 8 (1900–49) were intended to represent British samples only in the first instance, but things became asymmetric when American legal texts were added to those periods, too.[7] Besides, some fiction files carried an extra A in their extension, presumably to indicate ‘addenda’: whereas F1A and F6A were of British origin, F4A and F0A samples were American. When texts were added to ARCHER-2 the labelling became more opaque with ‘A’s for American texts (5A, 8A), ‘B’s for early British texts (0B), and some double figures for other new American samples (44, 77, 00).

A new, more transparent and systematic filename system was needed. ARCHER 3.1 achieved this by renaming all the files that make up this version of the corpus, old and new. Filenames now consist of 8+dot+3 characters, always in lower case, according to the formula nnnnabcd.gpv, where

- nnnn = year
- abcd = author abbreviation
- g = genre (a = advertising, d = drama, f = fiction, h = sermons, j = journal/diary, l = legal, m = medicine, n = news, s = science, x = letters)
- p = period (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)
- v = variety (a = American, b = British)

For example, the file 1650penn.j2b is a text written in 1650 by William Penn, is a journal (j), belongs to the period 1650–99 (2), and is British (b). (The filename in earlier versions was 1650PENN.J1) It must be recalled that in ARCHER 3.1 advertising and legal texts were not included, and that period 1 (1600–49) is not sampled given the exclusion of ARCHER-2 drama and fiction files. Notice as well that the new labelling system expands to period 0 (pre-1600) and period 9 (post-2000), but these are yet not represented in any version of the corpus. Table 4 lays out the correspondence between the new extensions representing period and variety in ARCHER 3.1 in comparison with the earlier ARCHER-1 and ARCHER-2 versions:

Table 4: Period extensions in filenames: correspondence between ARCHER 3.1 and ARCHER-1 + ARCHER-2

Period	**ARCHER 3.1** **b(ritish), a(merican)** **all genres**	**ARCHER-1** **ARCHER-2 (new extensions)**
pre-1600 (0)	–	–
1600–49 (1)	– –	0B British (Fiction and Drama)
1650–99 (2)	2b	1 British 1A British (Fiction)
1700–49 (3)	3b	2 British
1750–99 (4)	4b 4a	3 British 4 American 4A American (Fiction) 44 American (Advertising)
1800–49 (5)	5b 5a	5 British 5 American (Legal)* 5A American 50 American (Advertising)
1850–99 (6)	6b 6a	6 British 6A British (Fiction) 7 American 77 American (Advertising)
1900–49 (7)	7b 7a	8 British 8 American (Legal)* 8A American 80 American (Advertising)
1950–99 (8)	8b 8a	9 British 0 American 0A American (Fiction) 00 American (Advertising)
post-2000 (9)	–	–

In ARCHER 3.1, when the precise year of the sample text is unknown the formula is completed with x or xx (e.g. nnnxabcd.gpv or nnxxabcd.gpv), and the value of p in the extension will tell us the subperiod to which it belongs; for instance, the file 17xxarch.h4b is a sermon of unknown date but written within period 4 (1750–99).

The author abbreviation usually consists of the first four letters of the surname, padded out with hyphens if too short (e.g. 1697pix-.d2b for Mary Pix). Anonymous files may be abbreviated with 'anon' (e.g. 1735anon.m3b) or with a shortened form from the title (e.g. 1653merc.n2b from the newspaper *Mercurius Politicus*). When there is more than one author or more than one sample in the same file, the abbreviation is usually taken from the first author/sample, e.g. 1985dupo.m8a is a scientific text written by William D. Dupont and David L. Page. The same author should always have the same four-character abbreviation across sample texts, genres and periods (e.g. 1730fiel.d3b and 1743fiel.f3b for Henry Fielding's drama and fiction samples, respectively). However, when there is more than one sample from the same year the fourth character becomes a numeral, as in 1666cavn.x2b and 1666cav2.f2b for Mary Cavendish's letter and fiction samples. Note, however, that the reverse implication does not hold: the abbreviation 'hart', for instance, represents numerous different authors, e.g. Bret Harte in 1894hart.x6a, H. Hart in 1872hart.j6b, Anonymous in 1957hart.l8a.

Approximately 20 files had their filename revised in order to comply with the conventions established in 2006. For instance, author abbreviations consisting of three characters had '-' added to meet the standard four-digit formula (1675ray-.s2b), and different abbreviations for the same author were amended in favour of one single abbreviation across files and genres (e.g. 1720defo became 1720dfoe for the author Daniel Defoe).[8] A complete file list, including mapping to/from filenames used in ARCHER-1 and ARCHER-2, sorted or searched by year, author, genre and/or variety, can be obtained at the consortium universities and from the website; a sample is presented in Table 5:

Table 5: Sample of file list for ARCHER 3.1

year	auth	ext	g	p	v		old-year	old-auth	old-ext
1650	penn	j2b	j	2	b		1650	PENN	J1
1651	acon	x2b	x	2	b		1651	ACON	X1
1653	finc	x2b	x	2	b		1653	FINC	X1

4.2.4 Word counts

The header is enclosed in angle brackets (see Section 4.2.2). The second item in the header after the filename is now an accurate word count of the actual text – that is, of everything in the file which is not enclosed in angle brackets. Speaker names in curly brackets – used in some drama texts – are not counted either. That means that no header material is counted, nor the comments from the compilers enclosed in angle brackets. It also means that stage directions in some texts are not counted. Hyphenated words are counted as one item, as are all items other than punctuation surrounded by white space.

A typical word count per individual text is 2,000–2,500 words: although shortish by modern corpus standards, it is close to the existing average in ARCHER and so will not make new genres unbalanced. The exceptions to these word limit are fiction, with some texts reaching over 5,000 words; letters, with texts as small as 108 words, and only ten or so above 800 words; and British medical prose, with approximately sixty per cent of the texts below 1,800 words. If a particular file contains two or more texts, the total word count is provided in the header at the start of the file; in some cases word counts for individual texts are also given.

The Appendix gives the overall new counts for British and American varieties separately, with subtotals for each period and each genre. The grand total of ARCHER 3.1 under the above conditions is 1,789,309 words. The PERL script used to count ‘words’ was produced by Sebastian Hoffmann (2006). A full list of every ‘word’ in the corpus and its frequency is also available from the ARCHER website.

4.2.5 Spelling and punctuation

Another matter which enhanced the corpus in its ARCHER 3.1 version was the consistent use of hyphens and dashes. All dashes appear now as a double hyphen with one space to both left and right – like that – unless at line-end or beginning. 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 are not used, and hyphens no longer 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 (see Sample 3):

<1864mack.m6b, 2764 words, was 1864MACK.M6 -- M6. MacKay, Alexander. Naval Contributions, no.
IV. Edinburgh Medical Journal, vol. 9. [1864#4 in Atkinson EMJ
corpus; later shortened]>

The next case
occurred on the 20th of January, in the person of Wm. Palmer,
aet. 52, a leading seaman. The attack commenced during the
middle-watch, when he was seized with nausea and vomiting. On
his presenting himself in the morning he had a peculiar, heavy,
depressed appearance. He was trembling, and he stated that he
had had rigors. There was severe frontal headache **-- the bowels**
had been somewhat relaxed -- the skin was hot -- the pulse
accelerated -- and there was anorexia, and much thirst.

Sample 3: Use of hyphens and dashes in ARCHER 3.1: extracts from a 19th-century drama text (1864mack.m6b)

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*). Wherever possible the present version of ARCHER 3.1 displays such characters correctly if the text encoding is set to Western (Windows Latin 1). A list of all such characters was drawn up so that users who find them displaying wrongly in a text can use that list to interpret them or even change them appropriately on their own systems. There is a version of the ARCHER 3.1 files where such characters are stored not in Windows form but as HTML entities (e.g. *&auml; &pound; &deg;*), which will make it easier in the future to move ARCHER towards XML or Unicode. The default version of ARCHER 3.1 uses bracketed pseudo-words instead of the two mathematical operators which look like angle 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: *&lt; &gt; &alpha; &beta;* etc.

4.3 Access to the materials

4.3.1 Consortium members

ARCHER has its roots in the American universities of Northern Arizona and Southern California. The consortium of participant universities has naturally expanded over the years so that a total of ten universities in five countries were involved in the production of ARCHER 3.1 in 2006. As from autumn 2009, the consortium working towards the production of ARCHER 3.2 consists of fourteen universities in seven different countries. These are (by country): Northern Arizona, Southern California, Michigan (USA); Helsinki (Finland); Uppsala (Sweden); Bamberg, Freiburg, Heidelberg, Trier (Germany); Zurich (Switzerland); Lancaster, Manchester, Salford (UK); Santiago de Compostela (Spain). Manchester has been co-ordinating all project activities since December 2008.

At present, the corpus can only be accessed on-site at the consortium universities and under the conditions laid out in the user agreement form. ARCHER, in any of its versions, shall only be used for non-profit research and may not be distributed, or transferred to a third party in any format. Publications making use of the ARCHER-3.1 corpus shall include a reference to the name of the corpus, the years of compilation, and the compiler team, as set forth below (as in the latest agreement form, October 2009, available on the website):

> 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 University, University of Southern California, University of Freiburg, University of Heidelberg, University of Helsinki, Uppsala University, University of Michigan, University of Manchester, Lancaster University, University of Bamberg, University of Zurich, University of Trier, University of Salford, and University of Santiago de Compostela.

4.3.2 Format

ARCHER 3.1 exists in several slightly different versions on the same CD to suit different users and uses:

- The basic version with 955 separate text files, in the Windows Latin 1 character set, placed in a single folder.
- For convenience, a single ZIP-archive of exactly the same files, unzipped, but with each genre-period-variety combination in a separate folder (80 in total). Files can be extracted with or without its folder structure intact.

- 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.
- An alternative version with only 7-bit ASCII characters used, containing HTML-like entities such as *&auml; &eacute; &pound;* 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 – 169 files differ – are stored in a ZIP archive.

All files in ARCHER 3.1 have been date-stamped as 31 July 2006 03:10 to allow subsequent changes to show up easily in directory listings. The texts can be opened and read with any text editor, e.g. Microsoft Word, Wordpad, Notepad, Notepad++, etc., and they can be easily searched with concordance programs such as MonoConc, Wordsmith.

ARCHER 3.1 is neither tagged nor parsed, which implies that searches must be based on words, or parts/combinations thereof. A tagged version of the ARCHER-1 corpus does exist and has been extensively used in Biber and others' multi-dimensional factor analyses. The texts were tagged for grammatical/functional categories in three stages: firstly, by pre-editing the samples to replace dialect and non-standard spellings; secondly, automatically on a DOS-based computer using a combination probabilistic/rule-based tagger developed at Northern Arizona University (based on large online dictionary data, probabilities derived from the tagged LOB and Brown Corpora, and idiom lists); and thirdly, by checking selected words against an interactive post-editor which includes ambiguous and non-dictionary entries (see Biber 2001: 94–95). This tagged version can be accessed on request, should anyone be interested in the old version files. It is intended that the forthcoming ARCHER 3.2 will be morphologically tagged with the CLAWS8 tagset.

4.3.3 Documentation

As mentioned earlier in this paper, one of the added values of the project resulting in ARCHER 3.1 was the thorough elaboration of the corpus documentation (since 2005), including:

- detailed tables displaying the number of files and words per period, genre and variety;
- a complete file list, including mapping to/from filenames used in previous versions, sortable by many criteria;
- a complete word list, with frequencies;
- a PERL script for counting words;

- a list of non-ASCII characters and how they are coded;
- the user agreement form.

At Manchester we maintain a website offering background information about the project stages, as well as a comprehensive bibliographic database with information for each and every file in ARCHER – earlier, current and forthcoming versions. The database exists in FileMaker Pro format, but data can be exported to Excel. The present design consists of 30 subfields, thematically grouped as indicated below.

1. Filename
   a. as in ARCHER 3.1, following the new protocol nnnnabcd.gpv
   b. old filename, as in ARCHER-1 and/or ARCHER-2
   c. old filename from the sample header, as in ARCHER-1 and/or ARCHER-2 (not always the same as (1b))
2. Genre
3. Date
   a. Period, 50-year division
   b. Year
4. Variety
5. Extension
   a. as in ARCHER 3.1, following the new protocol nnnnabcd.gpv
   b. old extension, as in the original ARCHER-1 and/or ARCHER-2
6. Author
   a. Name
   b. Gender
7. Source
   a. Title of the sample text, if any
   b. Full reference of the edition consulted
8. Header
   a. Header: further bibliographic information provided by the compiler
   b. Header extras: further annotations by the compiler, e.g. about footnotes
   c. Spelling: notes by the compiler concerning spelling and/or italics; also mark-up issues related to dialogue in drama and fiction; notes on dialect
9. Word counts
   a. as in ARCHER 3.1, from our PERL script (2006)
   For ARCHER-1 and ARCHER-2 text files only:
   b. as in 2005, from a PERL script run at Manchester in preparation for ARCHER 3.1
   c. as stated in the original sample texts or word lists, if any

10. Compilation
    a. Version
    b. Tagging
    c. Bibliographic list provided by the compiler, if any
    d. University
11. Notes
    a. September 2005, on files from ARCHER-1 and ARCHER-2
    b. April 2006, in preparation for ARCHER 3.1
    c. July 2006, ARCHER 3.1
    d. on ARCHER 3.2 (since 2009)
12. Other information, incomplete
    a. addressee's name
    b. addressee's gender
    c. participants' relationship, to date only for British letters (1650–1899)

5 *Outlook: Towards ARCHER 3.2*

At the time of writing this paper (October 2010), we are close to the end of the current three-year project period from January 2008 to December 2010. The output of this phase will be known as ARCHER 3.2, envisaged to be completed in 2011.

As with ARCHER 3.1, the rationale behind producing yet another version is to expand and enhance the corpus. As Rissanen (1989: 17) had pointed out, one way of avoiding the 'God's truth fallacy' is "to keep the corpus open-ended – to structure it in a way that makes improvement and supplementation easy and uncomplicated". Our main goals are:

(i) to improve the accuracy of the texts and the genre classification by bringing in more periods and genres, and filling in gaps, especially of American English, to make the two main national varieties more comparable;

(ii) to add morpho-syntactic tagging to allow more general linguistic queries than mere string searches.

ARCHER 3.2 will consist of all the files in ARCHER-3.1 plus subsequent materials compiled since 2006; it will also include those ARCHER-1 and ARCHER-2 files omitted from ARCHER 3.1. The diachronic and textual coverage of the corpus is gradually being expanded. The target size of the corpus is open and will depend on resources available to the parties involved. Our expectation is that ARCHER 3.2 will consist of eight periods, by pushing the time-span back to 1600–49, and eleven genres, by (a) rescuing the advertising genre from the American texts included in ARCHER-2 and completing the partially compiled British counterpart; (b) restoring legal texts from American files in ARCHER-1 and adding new British legal texts (cf. López-Couso and Méndez-Naya forthcoming); and (c) splitting the single category journals/diaries into two: journals (j) vs. diaries (y).

Universities outside the UK are contributing new texts to help to fill the gaps in coverage in national varieties, genres and periods. New materials already to hand are British and American sermons (1750–1999), American letters and journals (1800–1849), American science texts (1800–1849, 1900–1949), and British legal texts (1900–1999). This amounts to 112 brand new files and ca. 235,300 words: 55 from British English (ca. 116,000 words) and 57 from American English (ca. 119,400 words). There will shortly be added more British and American letters (1650–1799, 1750–1799, respectively), British advertising (1750–1999), British legal texts from earlier periods (1700–1899), and early British prose and drama (1600–1649). In addition, as hinted above, the so far combined genre journals/diaries will be re-organised as two separate genres, and existing texts will be assigned to the appropriate new categories. To date, British texts from 1650 to 1899 have already been dealt with.[9]

The three UK universities are working closely together on the revision and correction of files in preparation for a tagged version of the corpus. At Manchester we will improve the quality of the sample texts, since some of the older files contain uncorrected scanned text, while others have sporadic replacement of original spelling by modern spellings or even different words, with the original text in brackets. Newly submitted texts will be checked for compatibility with the coding guidelines, while the oldest, unreliable texts will be brought up to the standard of the rest of the corpus, for instance, by checking locally unavailable texts at the British Library. Lancaster and Salford will carry out the automatic phase of the tagging of the entire corpus, both existing texts and new additions. The tagger, CLAWS8, is broadly compatible with that used in the *British National Corpus*. The automated tagging and sentence boundary marking will be subjected to systematic manual post-editing to correct the inevitable small percentage of errors. A master-copy of the tagged corpus will be stored in a form which could easily be migrated to XML in the future. Lastly, ARCHER is essentially an original-spelling corpus, albeit mostly based on editions; a search facility that allows use of modern-spelling equivalents may be added, too.

To conclude, although yet a smallish corpus by modern standards, we believe the improvements in ARCHER 3.1 and the present phase of development towards ARCHER 3.2 will greatly increase the corpus utility for the study of the history of the English language. The motive of the original compilers endures: "[i]t is our hope that ARCHER and corpora like it will foster the investigation of a wide array of questions that have remained uninvestigated or under-investigated in the past" (Biber *et al.* 1994a: 13).

Acknowledgements

Members of the ARCHER consortium would like to thank the various institutions which have financially or otherwise supported the development of ARCHER: National Science Foundation, Grant (BNS-9010893), Northern Arizona University / University of Southern California, 1990–93; *VARIENG*, the Centre of Excellence for the Study of Variation, Contacts and Change in English funded by the Academy of Finland at the University of Helsinki; Department of English at Uppsala University; British Academy Small Research Grant (SG-101087), The University of Manchester, 2010–11; Spanish Ministry for Science and Innovation and European Regional Development Fund (grant HUM2007-60706) and the Autonomous Government of Galicia (grants 2008-047 and INCITE-08PXIB204016PR), University of Santiago de Compostela.

Notes

[1] I am most grateful to David Denison for his helpful suggestions while writing this paper. Sections 4.2, 4.3 and 5 in particular have been adapted from earlier documentation of the corpus by David Denison, Sebastian Hoffmann and Nadja Nesselhauf (July 2006).

[2] As Biber *et al.* (1994a: 3, note 3) report, ARCHER was intended to comprise more written and speech-based registers, but, unfortunately, this was prevented by the lack of time and resources available then and later. The idea is not yet abandoned, though. For reference, the registers the original compilers had in mind are the following:

– written categories: official government documents, general prose, non-fictional narrative (biography, history, travelogue, gossip/social);

– speech-based categories: transcribed speech, monologue from public speeches, dialogues in parliamentary debate and town meetings, courtroom testimony.

[3] On the value of small size corpora, see, in particular, Nurmi (2002); Hundt and Leech (forthcoming).

4. A list of the different files in ARCHER-1a vs. ARCHER-1b can be obtained on request. The differences between ARCHER-1a and ARCHER-1 have not been tracked.
5. Although not included in ARCHER-2, British texts had originally been compiled in parallel to the American variety. The extant data consist of nearly 99,000 words (unevenly) distributed in 32 files for the time-span 1750–1999 (see further in Section 5).
6. Word counts for ARCHER-1 and ARCHER-2 are based on a PERL script run in 2005 at the University of Manchester, slightly different from the PERL script used for ARCHER 3.1 in 2006.
7. According to Biber *et al.* (1994a: 4, note 4), legal texts in periods 5 (1800–49) and 8 (1900–49) were to be labelled with an extra 'A' in the extension (5A and 8A) in order to indicate their American origin, as these periods originally comprised British files only. However, the file extension of these files in the version of ARCHER-1 available at Manchester in 2005 was simply 5 and 8. This is indicated with an * in Table 4.
8. Unfortunately, some filenames escaped revision at the time ARCHER 3.1 was completed and were only identified in subsequent revisions. These will be revised in the next version, ARCHER 3.2, and the errata will be duly made available.
9. Yáñez-Bouza (in prep.) will provide a detailed account of this split: criteria, procedure and contents of ARCHER before and after the redistribution of texts.

References

ARCHER website. http://llc.stage.manchester.ac.uk/research/projects/archer/.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.

Biber, Douglas. 2001. Dimensions of variation among eighteenth-century speech-based and written registers. In H-J. Diller and M. Görlach (eds.). *Towards a history of English as a history of genres*, 89–109. Heidelberg: Universitätsverlag C. Winter.

Biber, Douglas. 2004. Modal use across registers and time. In A. Curzan and K. Emmons (eds.). *Studies in the history of the English language II. Unfolding conversations*, 188–216. Berlin and New York: Mouton de Gruyter.

Biber, Douglas, Susan Conrad and Randi Reppen. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Biber, Douglas and Edward Finegan. 1988. Drift in three English genres from the eighteenth to the twentieth centuries: A multidimensional approach. In M. Kytö, O. Ihalainen and M. Rissanen (eds.). *Corpus linguistics, hard and soft. Proceedings of the 8th International Conference on English Language Research on Computerized Corpora*, 83–101. Amsterdam: Rodopi.

Biber, Douglas and Edward Finegan. 1989. Drift and the evolution of English style: A history of three genres. *Language* 65 (3): 487–517.

Biber, Douglas and Edward Finegan. 1992. The linguistic evolution of five written and speech-based genres from the seventeenth to the twentieth centuries. In M. Rissanen, O. Ihalainen, T. Nevalainen and I. Taavitsainen (eds.). *History of Englishes: New methods and interpretations in historical linguistics*, 688–704. Berlin and New York: Mouton de Gruyter.

Biber, Douglas and Edward Finegan. 1997. Diachronic relations among speech-based and written registers in English. In T. Nevalainen and L. Kahlas-Tarkka (eds.). *To explain the present: Studies in the changing English language in honour of Matti Rissanen*, 253–275. Helsinki: Société Néophilologique.

Biber, Douglas, Edward Finegan and Dwight Atkinson. 1994a. *ARCHER* and its challenges: Compiling and exploring *A Representative Corpus of Historical English Registers*. In U. Fries, P. Schneider and G. Tottie (eds.). *Creating and using English language corpora. Papers from the 14th International Conference on English Language Research on Computerized Corpora, Zurich 1993*, 1–13. Amsterdam: Rodopi.

Biber, Douglas, Edward Finegan, Dwight Atkinson, Ann Beck, Dennis Burges and Jena Burges. 1994b. The design and analysis of the *ARCHER* corpus: A progress report [*A Representative Corpus of Historical English Registers*]. In M. Kytö, M. Rissanen and S. Wright (eds.). *Corpora across the centuries: Proceedings of the First International Colloquium on English Diachronic Corpora, St Catharine's College Cambridge, 25–27 March 1993*, 3–6. Amsterdam and Atlanta: Rodopi.

Finegan, Edward and Douglas Biber. 1995. *That* and *zero* complementisers in Late Modern English: Exploring *ARCHER* from 1650–1990. In B. Aarts and C. F. Meyer (eds.). *The verb in contemporary English*, 241–257. Cambridge: Cambridge University Press.

Hickey, Raymond, Merja Kytö, Ian Lancashire and Matti Rissanen (eds.). 1997. *Tracing the trail of time. Proceedings from the Second Diachronic Corpora*

Workshop. New College, University of Toronto, May 1995. Amsterdam and Atlanta: Rodopi.

Hundt, Marianne and Geoffrey Leech. Forthcoming. Small is beautiful: On the value of standard reference corpora for observing recent grammatical change. In E. Traugott and T. Nevalainen (eds.). *Handbook on the history of English: Rethinking approaches to the history of English*. Oxford: Oxford University Press.

Kytö, Merja and Matti Rissanen. 1993. General introduction. In M. Rissanen, M. Kytö and M. Palander-Collin (eds.). *Early English in the computer age: Explorations through the Helsinki Corpus*, 1–17. Berlin and New York: Mouton de Gruyter.

Kytö, Merja and Matti Rissanen. 1995. English historical corpora: Report on developments in 1993–94. *ICAME Journal* 19: 145–158.

Kytö, Merja, Matti Rissanen and Susan Wright (eds.). 1994. *Corpora across the centuries: Proceedings of the First International Colloquium on English Diachronic Corpora, St Catharine's College Cambridge, 25–27 March 1993*. Amsterdam and Atlanta GA: Rodopi.

López-Couso, María José and Belén Méndez-Naya. Forthcoming. Compiling British English legal texts: A contribution to ARCHER. In N. Vázquez (ed.). *Creation and use of historical English corpora in Spain*. Newcastle upon Tyne: Cambridge Scholars Publishing.

Nurmi, Arja. 2002. Does size matter? *The Corpus of Early English Correspondence* and its sampler. In H. Raumolin-Brunberg, M. Nevala, A. Nurmi and M. Rissanen (eds.). *Variation past and present. VARIENG studies on English for Terttu Nevalainen*, 173–184. Helsinki: Société Néophilologique.

Rissanen, Matti. 1989. Three problems connected with the use of diachronic corpora. *ICAME Journal* 13: 16–19.

Rissanen, Matti, Merja Kytö and Kirsi Heikkonen (eds.). 1997. *English in transition: Corpus-based studies in linguistic variation and genre styles*. Berlin and New York: Mouton de Gruyter.

Rissanen, Matti, Merja Kytö and Minna Palander-Collin (eds.). 1993. *Early English in the computer age: Explorations through the Helsinki Corpus*. Berlin and New York: Mouton de Gruyter.

Yáñez-Bouza, Nuria. In preparation. *Journals and diaries in ARCHER – Revisited*.

Appendix

ARCHER-1 number of files and words in each category (version ARCHER-1b)

ARCHER-1 British

		A_{dverts}	D_{rama}	F_{iction}	H_{sermon}	$J_{our\&Diary}$	L_{egal}	$M_{edicine}$	N_{ews}	S_{cience}	$X_{letters}$	**Totals**
1600-49	1											
1650-99	2		*10*	*11*	*5*	*10*		*10*	*11*	*10*	*25*	***92***
			29,149	37,274	11,152	21,393		7,526	24,430	20,359	12,681	**163,964**
1700-49	3		*9*	*11*	*5*	*10*		*10*	*10*	*10*	*37*	***102***
			22,395	43,584	10,669	21,464		16,754	21,631	20,785	16,417	**173,699**
1750-99	4		*9*	*11*	*5*	*10*		*10*	*10*	*10*	*27*	***92***
			23,915	50,069	11,078	21,860		6,603	25,867	20,577	12,496	**172,465**
1800-49	5		*10*	*12*	*5*	*10*		*10*	*10*	*10*	*26*	***93***
			29,270	56,835	11,097	21,776		26,101	22,790	20,992	14,056	**202,917**
1850-99	6		*10*	*11*	*5*	*10*		*10*	*10*	*10*	*26*	***92***
			33,054	48,222	10,959	22,706		31,931	23,029	21,748	10,732	**202,381**
1900-49	7		*11*	*11*	*5*	*10*		*10*	*10*	*10*	*29*	***96***
			26,921	52,318	10,579	22,084		20,214	21,914	21,380	12,410	**187,820**
1950-90	8		*11*	*13*	*5*	*10*		*10*	*10*	*10*	*28*	***97***
			27,609	61,485	10,198	22,248		18,198	24,041	21,353	11,292	**196,424**
T files			***70***	***80***	***35***	***70***		***70***	***71***	***70***	***198***	***664***
T words			**192,313**	**349,787**	**75,732**	**153,531**		**127,327**	**163,702**	**147,194**	**90,084**	**1,299,670**

ARCHER-1 American

		A_{dverts}	D_{rama}	F_{iction}	H_{sermon}	$J_{our\&Diary}$	L_{egal}	$M_{edicine}$	N_{ews}	S_{cience}	$X_{letters}$	**Totals**
1600-49	1											
1650-99	2											
1700-49	3											
1750-99	4		*5*	*11*	*5*	*10*	*12*		*10*		*31*	***84***
			15,982	42,424	10,994	22,119	39,396		22,230		14,058	**167,203**
1800-49	5						*12*					***12***
							32,987					**32,987**
1850-99	6		*10*	*10*	*5*	*10*	*10*	*10*	*10*		*28*	***93***
			26,266	39,505	10,750	22,551	28,362	20,438	21,887		11,239	**180,998**
1900-49	7						*11*					***11***
							21,442					**21,442**
1950-90	8		*10*	*10*	*5*	*10*	*12*	*10*	*10*		*31*	***98***
			26,771	44,058	10,430	22,157	25,319	22,559	26,044		14,073	**191,411**
T files			***25***	***31***	***15***	***30***	***57***	***20***	***30***		***90***	***298***
T words			**60,019**	**125,987**	**32,174**	**66,827**	**147,506**	**42,997**	**70,161**		**39,370**	**594,041**

ARCHER-1b, Grand total files = 962
ARCHER-1b, Grand total words = 1,893,711

ARCHER-2 number of *new* files and words in each category

ARCHER-2 British

		A_{dverts}	D_{rama}	F_{iction}	H_{sermon}	$\text{J}_{\text{our\&Diary}}$	L_{egal}	$\text{M}_{\text{edicine}}$	N_{ews}	S_{cience}	$\text{X}_{\text{letters}}$	**Totals**
1600-49	1		*10*	*10*								***20***
			30,551	32,723								**63,274**
T files			***10***	***10***								***20***
T words			**30,551**	**32,723**								**63,274**

ARCHER-2 American

		A_{dverts}	D_{rama}	F_{iction}	H_{sermon}	$\text{J}_{\text{our\&Diary}}$	L_{egal}	$\text{M}_{\text{edicine}}$	N_{ews}	S_{cience}	$\text{X}_{\text{letters}}$	**Totals**
1750-99	4	*3*										***3***
		9,125										**9,125**
1800-49	5	*1*	*10*	*10*					*10*			***31***
		2,821	39,218	44,681					37,084			**123,804**
1850-99	6	*8*										***8***
		24,481										**24,481**
1900-49	7	*10*	*10*	*10*					*10*			***40***
		30,461	44,503	53,399					15,482			**143,845**
1950-90	8	*10*										***10***
		29,567										**29,567**
T files		***32***	***20***	***20***					***20***			***92***
T words		**96,455**	**83,721**	**98,080**					**52,566**			**330,822**

ARCHER-2, Grand total files = 112
ARCHER-2, Grand total words = 394,096

ARCHER 3.1 number of files and words in each category

ARCHER 3.1 British

		A_{dverts}	D_{rama}	F_{iction}	H_{sermon}	$J_{\text{our\&Diary}}$	L_{egal}	M_{edicine}	N_{ews}	S_{cience}	X_{letters}	**Totals**
1600-49	1											
1650-99	2		*10*	*11*	*5*	*10*		*21*	*10*	*10*	*25*	***102***
			26,648	41,512	11,146	21,374		23,117	22,292	21,441	12,659	**180,189**
1700-49	3		*10*	*11*	*5*	*10*		*14*	*10*	*10*	*28*	***98***
			25,177	44,021	10,664	21,443		21,936	21,612	20,780	12,093	**177,726**
1750-99	4		*10*	*10*	*5*	*10*		*20*	*10*	*10*	*26*	***101***
			23,962	45,056	11,068	21,843		21,003	23,087	20,565	12,091	**178,675**
1800-49	5		*10*	*10*	*5*	*10*		*10*	*10*	*10*	*25*	***90***
			26,267	44,946	11,089	21,740		20,278	22,903	20,994	12,576	**180,793**
1850-99	6		*10*	*10*	*5*	*10*		*10*	*10*	*10*	*26*	***91***
			26,469	43,289	10,953	22,686		22,143	23,066	21,715	10,705	**181,026**
1900-49	7		*11*	*10*	*5*	*10*		*10*	*10*	*10*	*29*	***95***
			23,048	45,274	10,569	22,066		20,204	21,975	21,337	12,434	**176,907**
1950-99	8		*11*	*10*	*5*	*10*		*13*	*10*	*10*	*28*	***97***
			24,450	45,095	10,190	22,225		20,794	22,920	21,308	11,259	**178,241**
T files			***72***	***72***	***35***	***70***		***98***	***70***	***70***	***187***	***674***
T words			**176,021**	**309,193**	**75,679**	**153,377**		**149,475**	**157,855**	**148,140**	**83,817**	**1,253,557**

ARCHER 3.1 American

		$\mathrm{A}_{\mathrm{dverts}}$	$\mathrm{D}_{\mathrm{rama}}$	$\mathrm{F}_{\mathrm{iction}}$	$\mathrm{H}_{\mathrm{sermon}}$	$\mathrm{J}_{\mathrm{our\&Diary}}$	$\mathrm{L}_{\mathrm{egal}}$	$\mathrm{M}_{\mathrm{edicine}}$	$\mathrm{N}_{\mathrm{ews}}$	$\mathrm{S}_{\mathrm{cience}}$	$\mathrm{X}_{\mathrm{letters}}$	**Totals**
1600-49	1											
1650-99	2											
1700-49	3											
1750-99	4		*10*	*11*	*5*	*10*		*9*	*10*	*10*	*27*	***92***
			27,331	42,417	10,987	22,109		23,433	22,271	20,664	11,056	**180,268**
1800-49	5											
1850-99	6		*10*	*11*	*5*	*10*		*10*	*10*	*10*	*28*	***94***
			24,214	44,224	10,740	22,534		20,424	21,992	21,326	11,253	**176,707**
1900-49	7											
1950-99	8		*10*	*10*	*5*	*10*		*10*	*10*	*10*	*30*	***95***
			23,810	44,214	10,123	22,131		22,473	23,072	21,343	11,611	**178,777**
T files			***30***	***32***	***15***	***30***		***29***	***30***	***30***	***85***	***281***
T words			**75,355**	**130,855**	**31,850**	**66,774**		**66,330**	**67335**	**63,333**	**33,920**	**535,752**

ARCHER 3.1, Grand total files = 955
ARCHER 3.1, Grand total words = 1,789,309 words