

CLLC Histories

The CLLC "Histories Corpus" consists of a group of 115 authorial texts ranging from Defoe (1660-1731) through to authors born in the latter half of the twentieth century. Each author is represented by at least three texts, some by as many as eight or nine. The texts are all retrospective fictional narratives, couched in the first person and treating of the narrator's supposed experiences or observations. ("My name is "X" and this is the story of my life".) Most (but not all) of the narratives are embedded in larger works of fiction; however the texts are marked up so that the narrator's own words can be counted separately. This is achieved by the use of a fixed panel at the start of each line. The direct speech of the narrator of the history is always labelled [A] whilst other actual speakers are labelled [B], [C] and so on, with [V] being used to indicate any "minor female" speaker and [W] any "minor male" speaker. Reported speech is also distinguished by the use of lower case notation: eg. @a indicated the reported speech of the historian, @b, @c, @v & @w etc. being used for other reported speech. If there is another narrator, apart from the historian/story-teller, it is distinguished by the use of []. In those instances where the speaker is unidentifiable or where there is combined speech, [X] & [Y] are used. [Z] is reserved for instances of quoted text. This system was applied to all the texts of the 115 authors, making it possible to compare main speakers, minor speakers etc. of any group of texts across the range of over three centuries, male/female authorship and three different nationalities (British, American and Australian).

At the start of each authorial set, bibliographical information is given for each text selection. Each new text in the authorial file can be found by searching for @@@@. At the start of each new text a complete word count & character reference guide has been provided for the speakers of that text.

Since the most common words were used for the initial analyses, a number of these most common words were "homographed." That is, they were marked so that the computer could distinguish between their various grammatical forms and count these as separate words: - the most obvious example being "to" as an infinitive (marked to\$9\$) and "to" as a preposition (marked to\$4\$). The corpus is available with and without the use of homographs.

An XLS file with 4 worksheets accompany the histories:

- DateGroups: provides information concerning one of a number of chronological groups which were used at various times
- Wordtotals: provides information concerning the total number of words for each text selection and the number of [A] words used by the "historian"/storyteller
- Nationality: provides information concerning the nationality of the authors
- AuthorInfo: provides information concerning the birth date and nationality of the authors

Publications

A number of published articles have appeared over the years which have made use of various parts of the histories corpus. A list of these is appended.

- Quotes from a couple of these articles, where John Burrows describes what he had in mind as he developed the corpus of authorial histories.

From "Computers and the Study of Literature", John Burrows

"The corpus of texts amounts to 1.1 million word-tokens by 45 authors ranging from Defoe to the present day. Every author is represented by at least three 'texts': the smallest authorial sets, those of Dickens and Lodge, comprise around 8500 word-tokens; the largest, those of Henry and Sarah Fielding, comprise around 75000 word-tokens. The texts are all retrospective fictional narratives, couched in the first person and treating of the narrator's supposed experiences or observations. Only the narrator's own words are incorporated in the analysis. Most of the narratives are embedded in larger works of fiction: but in some of the more recent texts, the outer narrative frame is only lightly sketched and in a few texts, like Defoe's, there is none."

"My evidence suggests that, while the 'history' lay dormant, the language of English fiction continued its passage towards modern vernacularity. When the first-person narrative revives, there is some evidence of a subtle change in genre. ... my impression as a reader [is] that, in its modern form, the 'history' no longer treats so much of 'what happened to me' as of 'what I saw'. 'What I saw' is more likely to treat of things observed than 'what happened to me', which so often, in the earlier texts, emphasises 'what he did to me'. In the latter part of the nineteenth century, 'what I saw' commonly deals in experiences of the supernatural and the fantastic. In the twentieth century first-person narrative seems to have re-entered the literary mainstream. Very often, however, the new 'histories' present their own narrators ironically and explicitly project themselves as fictive."

From "Tiptoeing into the Infinite: Testing for Evidence of National Differences in the Language of English Narrative", John Burrows

"The texts chosen to represent each author are all retrospective narratives couched in the first person – a literary form known in the eighteenth century as the 'history.' Although most of them are embedded in larger works, there is no outer narrative framework in Defoe's novels or in some of the short stories. Where the outer narrative is also couched in the first person, outer and inner narratives do not often show marked stylistic differences. Most of the narratives purport to be spoken, whether to the reader or to another character. To diminish the effect of differences within their work, each author is represented by at least three narratives. Counting only the words of the first-person narrators and excluding all such interpolations as the conversations they report, most of the authorial sets range from 10,000 to 75,000 word-tokens."

A number of published articles have appeared over the years which have made use of various parts of the histories corpus. A list of these is appended.

Select Bibliography

Articles which made use of the Histories Corpus

Burrows, John, 2004: *Textual Analysis. A Companion to the Digital Humanities*. Edd. Susan Schreibman et al. Blackwell, Oxford.

Burrows, John, 1999: *Computers and the Idea of Authorship*. Reprinted by invitation in Fotis Jannidis. *Rückkehr des Autors. Zur Erneuerung eines umstrittenen Begriffs*. Edd. Gerhard Lauer, et al. Niemeyer Verlag, Tübingen, 133-44.

- Burrows, John, 1996: Tiptoeing into the Infinite: Testing for Evidence of National Differences in the Language of English Narrative. *Research in Humanities Computing '92*. Edd. Susan Hockey and Nancy Ide. Clarendon, Oxford, 1-33.
- Burrows, John, 1995: Computers and the idea of Authorship. *The Humanities and a Creative Nation: Jubilee Essays*. Ed. Deryck Schreuder, Australian Academy of the Humanities, Canberra, 89-108.
- Burrow, John, 1992: Not unless you ask nicely: The Interpretative Nexus between Analysis and Information, *Literary and Linguistic computing*, vii, 91-110.
- Burrow, John, 1992: Fossicking about the Territory: Testing for Specimens of an Australian Narrative Dialect, edd. Margaret Harris & Elizabeth Webby, *Reconnoitres: Essays in Honour of G.A. Wilkes*, Sydney University Press & Oxford University Press, 36-53, 241-9.
- Burrows, John, 1992: Computers and the Study of Literature, ed. Christopher Butler, *Computer and Written Texts: an Applied Perspective*, Blackwell, Oxford, 167-204.
- Burrow, John, 1991: I Lisp'd in Numbers: Fielding, Richardson and the Appraisal of Statistical Evidence, *The Scriblerian*, xxxiii, 234-41.
- Burrows, John, 1989: An Ocean where each Kind ...: Statistical analysis and some Major Determinants of Literary Style, *Computers and the Humanities*, xxiii, 309-21.
- Burrows, John & A.J. Hassall, 1988: "Anna Boleyn" and the Authenticity of Fielding's Feminine Narratives, *Eighteenth-Century Studies*, xxi, 427-53.